



US009449017B2

(12) **United States Patent**  
**Pitts**

(10) **Patent No.:** **US 9,449,017 B2**  
(45) **Date of Patent:** **Sep. 20, 2016**

(54) **DISTRIBUTED FILE SYSTEM  
CONSISTENCY MECHANISM EXTENSION  
FOR ENABLING INTERNET VIDEO  
BROADCASTING**

(71) Applicant: **William M Pitts**, Los Altos, CA (US)

(72) Inventor: **William M Pitts**, Los Altos, CA (US)

(\*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 16 days.

(21) Appl. No.: **13/959,508**

(22) Filed: **Aug. 5, 2013**

(65) **Prior Publication Data**

US 2013/0326572 A1 Dec. 5, 2013

**Related U.S. Application Data**

(60) Division of application No. 12/632,301, filed on Dec. 7, 2009, now Pat. No. 8,504,597, which is a continuation-in-part of application No. 12/626,116, filed on Nov. 25, 2009, now abandoned, which is a continuation-in-part of application No. 11/223,572, filed on Sep. 9, 2005, now Pat. No. 8,504,565.

(51) **Int. Cl.**

**G06F 17/30** (2006.01)  
**H04N 21/222** (2011.01)  
**H04N 21/61** (2011.01)  
**H04N 21/84** (2011.01)  
**H04N 21/654** (2011.01)  
**H04H 20/82** (2008.01)

(52) **U.S. Cl.**

CPC ... **G06F 17/30194** (2013.01); **G06F 17/30144** (2013.01); **H04N 21/222** (2013.01); **H04N 21/6125** (2013.01); **H04N 21/654** (2013.01); **H04N 21/84** (2013.01); **H04H 20/82** (2013.01)

(58) **Field of Classification Search**

None

See application file for complete search history.

(56) **References Cited**

**U.S. PATENT DOCUMENTS**

6,101,547 A \* 8/2000 Mukherjee et al. .... 709/231  
2010/0049757 A1 \* 2/2010 Thomas et al. .... 707/613

**OTHER PUBLICATIONS**

Javvin.com, Microsoft CIFS: Common Internet File System. Apr. 15, 2004.\*

\* cited by examiner

*Primary Examiner* — Syed Hasan

(74) *Attorney, Agent, or Firm* — Superior IP, PLLC; Dustin L. Call

(57) **ABSTRACT**

The consistency callback mechanisms employed by local file systems such as NTFS and distributed file systems such as DDS, NFS and CIFS are extended to provide a shared memory foundation for efficiently broadcasting real-time high definition video from a source object to large numbers of viewers via the Internet. Distributed applications such as video viewing client applications establish connections to a common distributed file system object, and then each application registers with the underlying distributed file system to receive notifications whenever the video source modifies the source object. The data required to update images maintained by viewing clients is included in notification messages. The distributed file system employs a network of proxy cache nodes. Proxy cache nodes receive notification messages (complete with image update data) and update their cached images of the source object and then retransmit the notification messages towards the viewing clients using IP multicast techniques. In this manner, the distributed file system's consistency mechanism efficiently employs network resources to enable the real-time distribution of video content streams.

**17 Claims, 5 Drawing Sheets**

Event
Event_Range_Start
Event_Range_End
Update_Range_Start
Update_Range_End
Notification_Routine_Address

Event:

```
#define ACCESS_READ 0x00000001
#define ACCESS_WRITE 0x00000002
#define ACCESS_RW 0x00000003
#define ACCESS_OPEN 0x00000004
#define ACCESS_CLOSE 0x00000008

#define SEND_UPDATE 0x00000100
```

Event Notification Descriptor

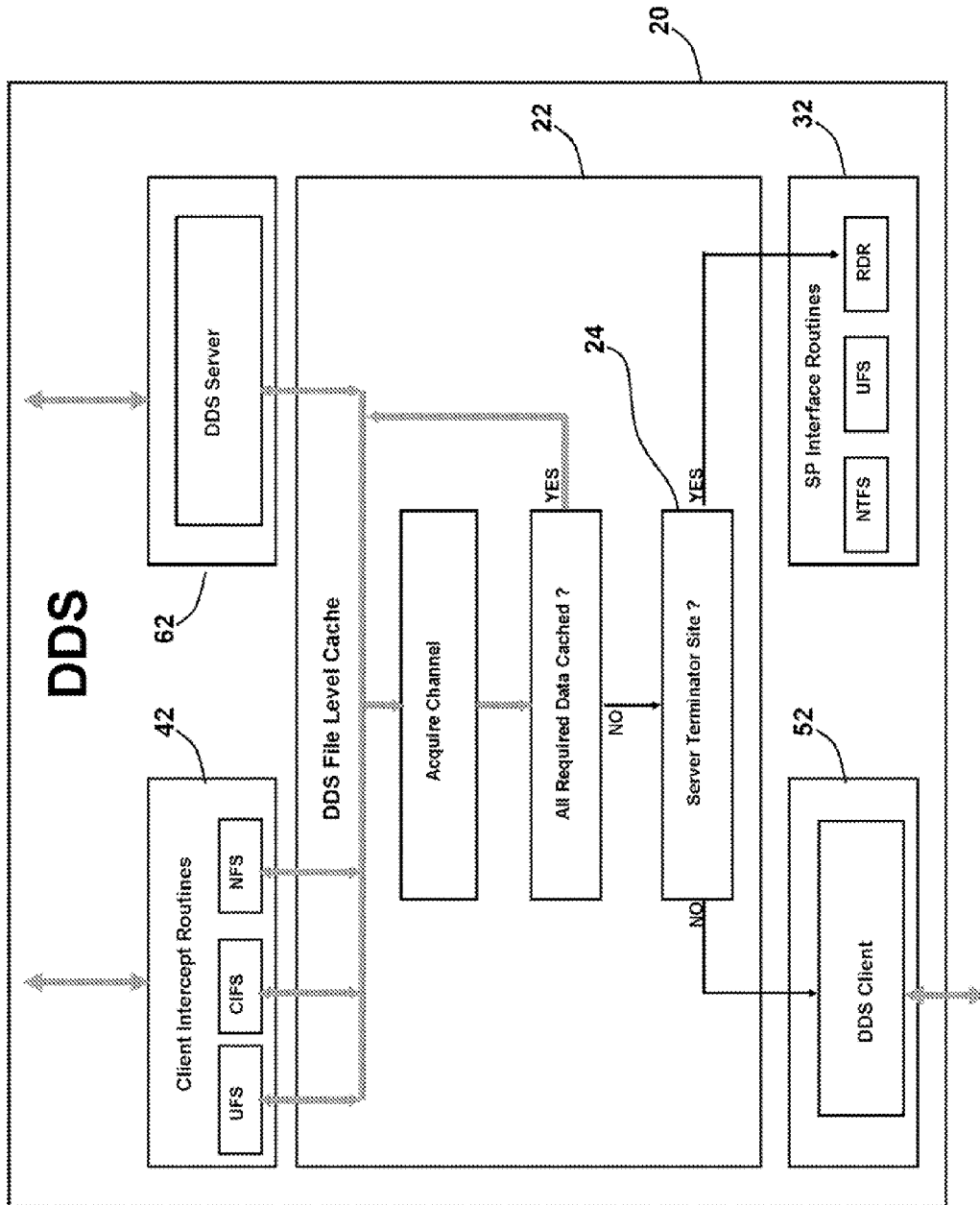


Fig. 1 Prior Art

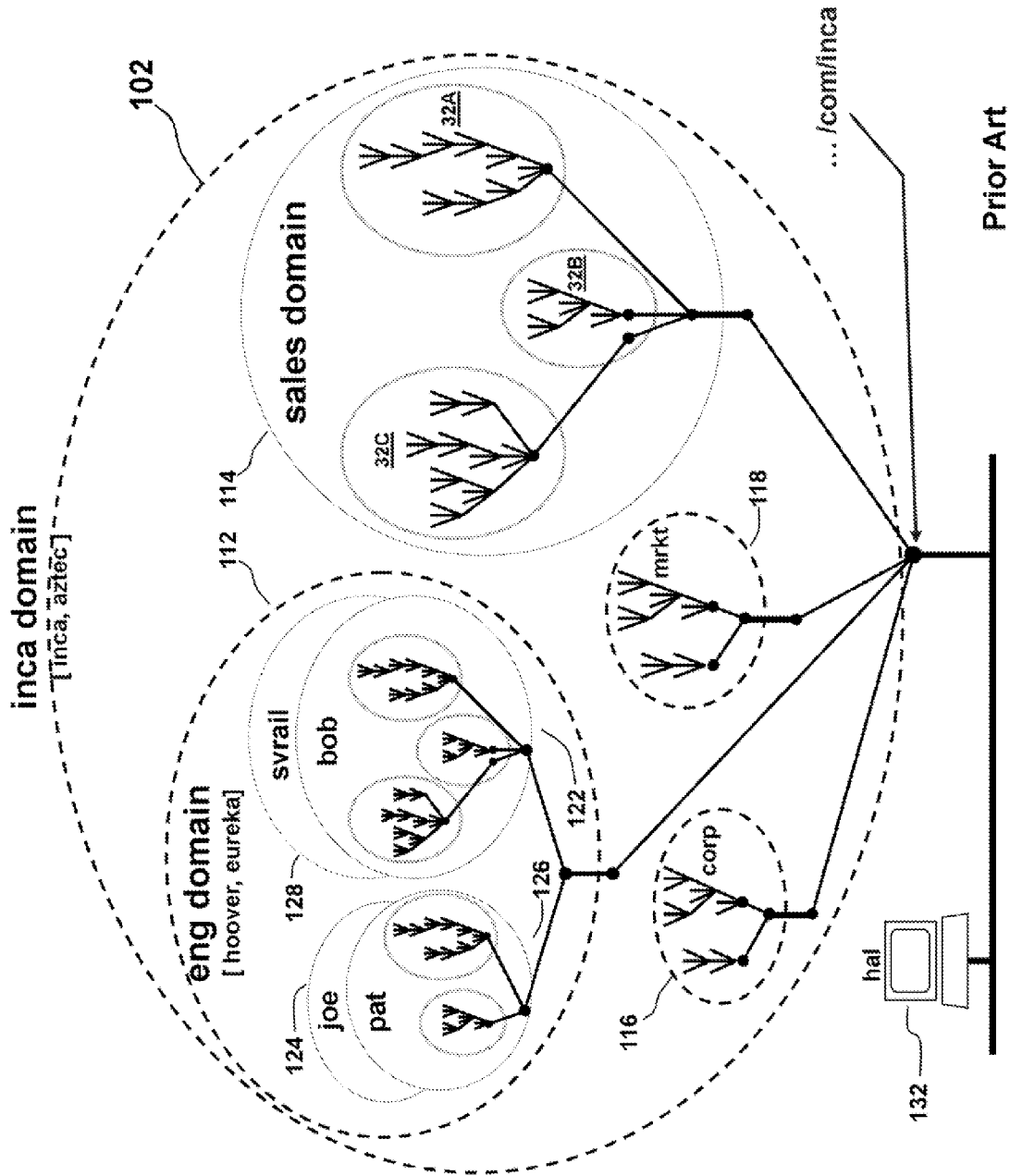


Fig. 2

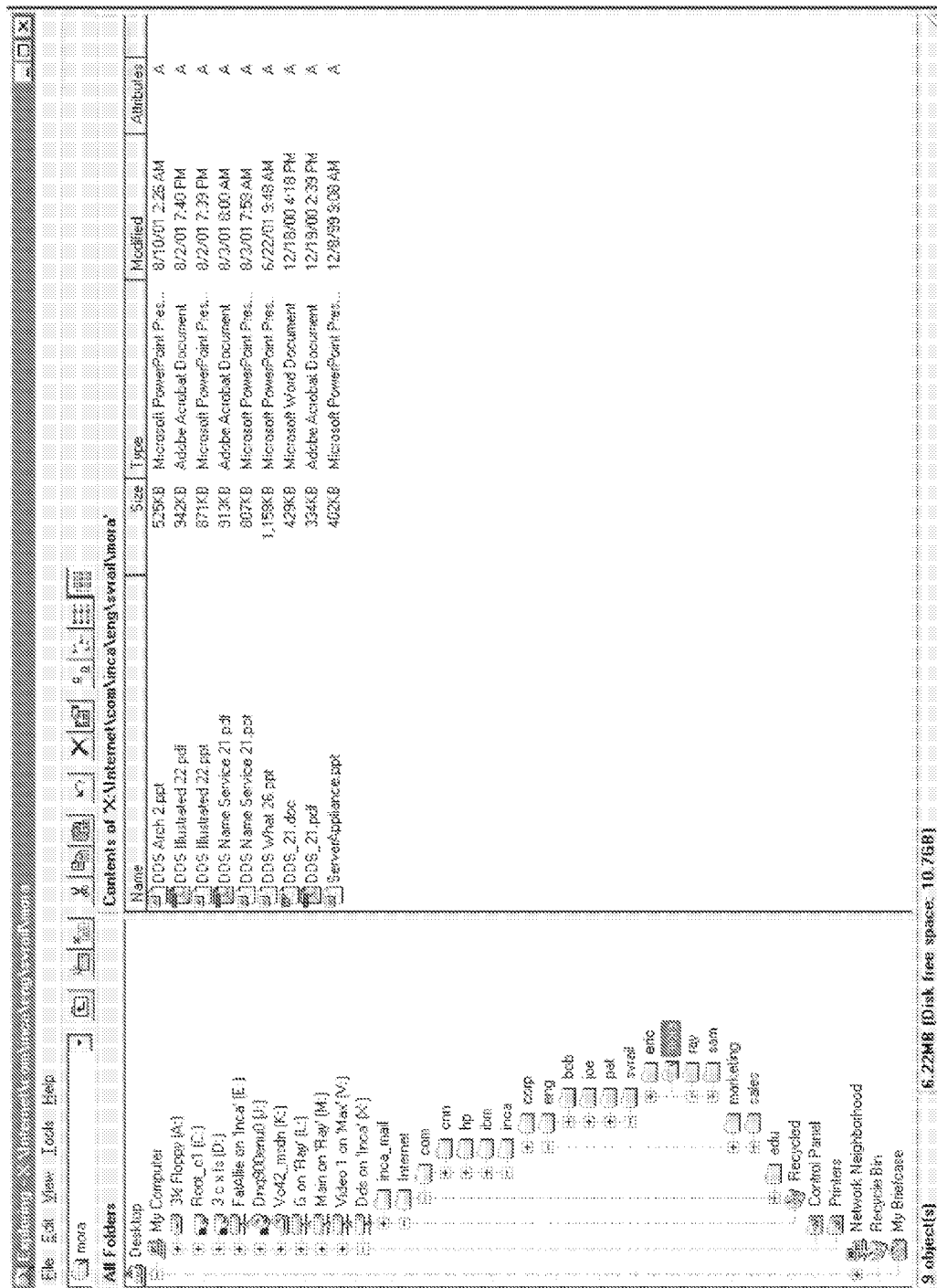


Fig. 3

Prior Art

Event
Event_Range_Start
Event_Range_End
Update_Range_Start
Update_Range_End
Notification_Routine_Address

Event:

#define	ACCESS_READ	0x00000001
#define	ACCESS_WRITE	0x00000002
#define	ACCESS_RW	0x00000003
#define	ACCESS_OPEN	0x00000004
#define	ACCESS_CLOSE	0x00000008
#define	SEND_UPDATE	0x00000100

Event Notification Descriptor

Fig. 4

Event
Event_Range_Start
Event_Range_End
New_Data_Start
New_Data_End
Site_Identifier

Event:

#define	FILE_WAS_READ	0x00000001
#define	FILE_WAS_WRITTEN	0x00000002
#define	FILE_WAS_OPENED	0x00000004
#define	FILE_WAS_CLOSED	0x00000008
#define	UPDATE_INCLUDED	0x00000100

Notification Message Descriptor

Fig. 5

# **DISTRIBUTED FILE SYSTEM CONSISTENCY MECHANISM EXTENSION FOR ENABLING INTERNET VIDEO BROADCASTING**

This application is a divisional application of, and claims the benefit of and priority to, U.S. patent application Ser. No. 12/632,301 filed on Dec. 7, 2009, which application is incorporated herein by reference in its entirety.

U.S. patent application Ser. No. 12/632,301 is a continuation-in-part of U.S. patent application Ser. No. 12/626,116, filed on Nov. 25, 2009, which is a continuation-in-part of U.S. patent application Ser. No. 11/223,572, filed on Sep. 9, 2005.

## **BACKGROUND**

### **1. Technical Field**

The present disclosure relates generally to the real-time broadcasting of audio and video streams to global audiences via the Internet.

U.S. patent application Ser. No. 12/626,116, of which the present application is a continuation-in-part, disclosed systems and methods for extending the functionality of distributed file system consistency callback mechanisms, enabling distributed applications, each having established a connection to a common object, to receive out-of-band notification messages whenever that object is accessed or modified.

This application further describes how the systems and methods disclosed in the parent application enable the real-time broadcasting of audio and video streams to global audiences via the Internet. The “broadcasting station” may be a desktop or laptop computer, or even a video camera equipped cellular telephone. This advancement in networking technology empowers ordinary individuals with the ability to directly communicate with vast numbers of people from virtually any location. Broadcasting enterprises such as NBC, ABC, CNN, Fox News, etc., will no longer be the gatekeepers of broadcasting. In the near future every individual will have the ability to address a global audience.

### **2. Description of Background Art**

The volume of information contained within a single file system has increased dramatically since file systems were first designed and implemented. Whereas early file systems managed tens of megabytes of data, today’s distributed file systems often encompass tens of terabytes. This represents a million fold increase, and the end is not in sight. Consider the following:

- a. The storage capacity of a 3.5" disk drive is projected to increase from today’s 250 gigabytes to 25 terabytes.
- b. A single file server typically exports the file systems resident on many disk drives.
- c. Global file systems, now being deployed, will wrap hundreds or thousands of file servers into a single virtual file server.

The volume of data contained within a file system will be so enormous that information may be lost within the file system! Current file systems provide only a very limited ability to locate particular information contained in the system’s files.

Over time, the file system application programming interface (“API”) of the earliest file system implementations has been enhanced to include new functionalities. And, it is time once again to extend the definition and capabilities of a file system.

The Distributed Data Service (“DDS”) is a distributed file system that integrates industry standard file servers (Unix,

Linux, Windows, Mac) into highly distributed multi-protocol virtual file servers of vast proportions. A single DDS virtual file server may encompass hundreds of petabytes. Fundamental concepts underlying a DDS virtual file server are disclosed in U.S. Pat. Nos. 5,611,049, 5,892,914, 6,026,452, 6,205,475, 6,366,952(B2), 6,505,241(B2) and 6,804,706(B2). All of the immediately preceding United States patents are hereby incorporated by reference as though fully set forth here.

DDS global file systems accessible via a DDS virtual file server encompass entities that might not normally be thought of as files, so when describing DDS global file systems the term object is often used to denote a superset class which includes what is conventionally identified as a file.

The object related definitions are:

1. Object—A named entity represented within a namespace to which a connection can be established for the purpose of reading or writing data. The most common type of object is a file, but other types include:
  - a. directories, domains, and other containers,
  - b. live video feeds,
  - c. application programs, and
  - d. shared memory.
2. Object system—A provider of objects. For example, a file system (a type of object system) contains a collection of files and it provides a service through which its content may be accessed.
3. Provider—A synonym for object system.
4. Namespace—A set of names in which all names are unique. All objects within an object system have at least one name, and the complete set of all names for all objects comprises the object system’s namespace.

DDS constructs virtual file servers from heterogeneous collections of industry standard file servers. A single DDS virtual file server provides a highly distributed file service, perhaps, incorporating as many as thousands of geographically dispersed file servers. As stated previously, DDS is also capable of providing remote access to objects other than files, such as live video feeds. Accordingly, the term “object” is generally used throughout this document to denote a file, a data stream, or some other data entity that might stretch the definition of “file”.

The DDS architecture provides a framework for highly distributed, multi-protocol file caching. FIG. 1 illustrates a basic structure for a DDS cache module referred to by the general reference character 20. The DDS cache module 20 may be installed on file servers, client workstations, and intermediate network nodes such as routers, switches, and dedicated file caching appliances.

DDS implements a file level cache 22 layered above all data sources. A data source is usually a file system, either local or remote, but it could be, for example, a real time data stream. When appropriate, each DDS cache module 20 automatically caches whatever data is being accessed through its file level cache 22 regardless of the source of data.

Individual file level caches 22, using both local RAM and local disk for data storage, may vary dramatically in size. Some DDS cache modules 20, perhaps within switches and routers, may implement only RAM based caching. Other DDS cache modules 20 in high capacity locations might be configured with 16 gigabytes of RAM and a terabyte or more of disk storage.

Although some current distributed file system implementations employ callback mechanisms to synchronously invalidate file images cached “just below” the multiple client

processes accessing a shared file, all processes remain unaware of the consistency operations. When a process reads a file, the response includes the most recently written data (the modification), and whatever consistency operations were required to ensure the currency of the response remains hidden from file system clients.

Sprite, CIFS, and NFSv4 each implement consistency callback mechanisms as described above. Therefore, since cache consistency is maintained through private communications between the server and the client components of these distributed file systems, it is impossible for one process to detect another process's modification of a shared file except by reading the file. Consequently, detecting shared file modifications when using distributed file systems such as Sprite, CIFS, and NFSv4 requires use of a polling loop that periodically reads the shared file.

All DDS cache modules **20** maintain the consistency of cached images via origin file server callbacks. Files which are in use or have been recently used are registered with the origin file server, and may receive a callback (at the onset of a concurrent write sharing condition) to invalidate or flush the cached file image. DDS incorporates a consistency disconnect-reconnect mechanism, described in U.S. Pat. No. 5,946,690 ("the '690 patent"), whereby a cached file image, including the file's metadata, may be disconnected from the origin file server and then, at a later time (weeks, months, years), reconnected and revalidated. This is an essential mechanism for implementing high capacity, long term (persistent) caches. The '690 patent is hereby incorporated by reference.

A DDS cache module **20**, illustrated in FIG. 1, includes five major components:

#### 1. File Level Cache **22**

The file level cache **22** consists of a large number of channels (10,000 to 100,000). Each channel is a data structure that contains (or has pointers to) consistent data/metadata images of a (usually) remote source object. Channels also contain data structures that track client access patterns, measure rates of consumption by clients and the rate of replenishment from the origin file server.

Channels are managed on a least recently used ("LRU") basis and are identified by an object ID. A simple hash mechanism allows an incoming file system request to be connected to the appropriate channel within a microsecond or two. Background processes strive to ensure that, for well-mannered clients, channels are primed such that incoming requests are immediately serviced with no need to block while waiting to fetch data from downstream (closer to the origin file server).

Just before a channel is reassigned to a new object, its contents are written to a disk cache if one exists at this DDS cache module **20**, not illustrated in FIG. 1. (Throughout this document 'this' in *italics* refers to a specific node or instance where the described action is occurring. In this case, the DDS cache module **20** where a channel is being reassigned to a new object.) Disk caches are also managed on an LRU basis.

The file level cache **22** also incorporates a redirector **24**. All cache misses are passed on to the redirector **24**, even when the source object resides within a local file system.

#### 2. Source Provider Interface Routines **32**

For any given file with multiple active clients distributed about the network, there exists a tree structured hierarchy of DDS cache modules **20** rooted at a DDS Server Terminator Site. The DDS Server Terminator Site communicates directly with the origin server to access a file. The Source Provider Interface Routines **32** ("SPIRs **32**") interface one or

more local file systems to the DDS cache module **20**, e.g. NTFS, UFS, RDR . . . . When a DDS cache module **20** is the DDS Server Terminator Site for a file, the file level cache **22** accesses the file via one of the SPIRs **32**.

#### 3. Client Intercept Routines **42**

A set of client intercept routines **42** provide industry standard local and remote file services directly to clients. The DDS cache module **20** with which a client communicates directly via one of the client intercept routines **42** is the DDS Client Terminator Site for that client. FIG. 1 depicts a DDS cache module **20** configured with three client intercept routines **42**: UFS, CIFS, and NFS. Unmodified Windows clients communicating directly with this DDS cache module **20**, for example, may use the CIFS protocol to access file data sourced from a Unix file server that is remote from this DDS cache module **20**, or for which this DDS cache module **20** is the file's DDS Server Terminator Site. Local processes running on the system hosting this DDS cache module **20** may access the same file data via the UFS (Unix File System) client intercept routine **42**.

Each file's metadata is represented within file level cache **22** as a discriminated union: the discriminator identifies 'UFS' as the source file system and the union contains the file's metadata as formatted by the UFS source provider routine on the file's origin server.

When a particular DDS cache module **20** services NFS or UFS requests, no protocol translation is necessary. However, the CIFS client intercept routine **42** must be configured with a UFS to CIFS translation module (not depicted in FIG. 1) so 'UFS' files may be accessed via the CIFS protocol.

Note that local clients may use the UFS interface to access remote files via the DDS cache module **20**.

#### 4. DDS Client Code **52**

When a file level cache **22** requires additional file data and the DDS cache module **20** is not the DDS Server Terminator Site for the file, the file level cache **22** invokes DDS Client code **52** to fetch missing file data. To access missing file data, the DDS Client code **52** generates and dispatches a network request expressed in a DDS protocol directed toward the file's DDS Server Terminator Site.

#### 5. DDS Server Code **62**

A DDS Server code **62** receives requests dispatched by the DDS Client code **52** at an upstream DDS cache module **20**, i.e. a DDS cache module **20** which is or is closer to the DDS Client Terminator Site. The DDS Server code **62** implements the DDS protocol.

#### DDS Protocol

The DDS protocol is a remote file access protocol, providing functionality comparable to NFS and/or CIFS. It is designed to efficiently stream file data and metadata into large RAM/disk caches distributed throughout a network, and to maintain the consistency of cached images at a level which approaches that of a local cache. The DDS protocol transfers and caches images of files and objects from many different sources (UFS, VxFS, NTFS file systems, video cameras, . . . ) with no "image degrading" translations between the source object and its cached image. Protocol translation is always performed at DDS Client Terminator Sites, and is required only for heterogeneous (with respect to the origin file server) clients.

The DDS protocol, as currently implemented, consists of five operations:

1. DDS\_CONNECT—This operation connects to an existing file, directory, or file system, and optionally creates a new file or directory if it doesn't already exist. If successful, this operation returns a file handle and the attributes (size, owner, date created, . . . ) of the target



## 5

- object. This operation supplies the functionality required by the NFS operations mount, lookup, create, and mkdir.
2. DDS\_NAME—This operation manipulates names in various ways. It supplies the functionality required by the NFS operations link, symlink, rename, remove, and rmdir.
  3. DDS\_LOAD—This operation loads data and metadata. The request includes flags (DDS\_CC\_SITE\_READING, DDS\_CC\_SITE\_WRITING) that inform downstream DDS cache modules 20 what types of operations will be performed upon data/metadata images cached at the DDS cache module 20. These flags are used by DDS's distributed consistency mechanism to keep track of the types of file activities occurring at various DDS cache modules 20.
- A single load or flush request may specify multiple file segments, and each segment may be up to 4 gigabytes in length.
- The response to a load or flush request includes flags (DDS\_CC\_SUSTAIN\_DIR\_PROJECTION, and DDS\_CC\_SUSTAIN\_FILE\_PROJECTION) that indicate whether the returned data/metadata may be cached or whether it must be discarded immediately after responding to the current client request.
- The DDS\_LOAD operation supplies the functionality required by the NFS operations statfs, getattr, setattr, read, write, readdir, and readlink.
4. DDS\_FLUSH—This operation flushes modified data downstream towards the DDS Server Terminator Site. A flush level specifies how far the flush should propagate. Currently available flush levels are:
    - a. DDS\_FLUSH\_TO\_NOWHERE—Don't flush
    - b. DDS\_FLUSH\_TO\_CCS—Flush to Consistency Control Site ("CCS")
    - c. DDS\_FLUSH\_TO\_SITE\_DISK—Flush to first DDS cache module 20 with disk cache
    - d. DDS\_FLUSH\_TO\_SITE\_STABLE\_RAM—Flush to first DDS cache module 20 with stable RAM
    - e. DDS\_FLUSH\_TO\_SERVER\_DISK—Flush all the way to origin server's disk
- A basic concept of DDS is that DDS projects the source file system at the DDS Server Terminator Site into distant DDS cache modules 20. Consequently, an image of data present in an upstream DDS cache buffer is identical to that in an internal file system buffer at the DDS Server Terminator Site. After a write operation modifies a file system buffer (either local or remote), performance is enhanced if the buffer is asynchronously written to the server's disk at the DDS Server Terminator Site. However, file modifications are safeguarded when they're synchronously written to disk or some other form of stable storage. Flush levels allow both the client and the DDS Server Terminator Site to express their level of paranoia regarding file consistency. The most paranoid of the client and the DDS Server Terminator Site prevails.
5. DDS\_FSCTL—DDS\_FSCTL implements various file system control operations. These various file system control operations include:
    - a. fs\_sync—Commands all downstream DDS cache modules 20 to flush all modified file data from this DDS cache module 20 and this file or file system to whatever level is specified by the flush level parameter.
    - b. fs\_ping—Pings for the status of specified file systems at downstream DDS cache modules 20. Usually,

## 6

ally, the fs\_ping request specifies all file systems currently being accessed through downstream DDS cache modules 20 regardless of the file system's origin server. Downstream DDS cache modules 20 respond immediately with status indications for each specified file system.

Upstream DDS cache modules 20 use fs\_ping (often referred to as a fast ping) to detect, within a few seconds, partitioning events that isolate DDS cache modules 20 from remote file systems. Fast ping rates (typically set from 500 to 3000 milliseconds) are specified as mount parameters when each file system is mounted. For a set of file systems accessed through the same downstream DDS cache module 20, the most aggressive rate determines the fast ping rate for that DDS cache module 20.

- c. fs\_callback—Pings the root of the specified file system at the next downstream DDS cache module 20. The downstream DDS cache module 20 doesn't respond until the timeout period (specified in the request, typically 5 to 30 minutes) expires or a consistency event occurs (on any file in the specified file system). Occurrence of a consistency event requires that a cached file image at the upstream DDS cache module 20 (and DDS cache modules 20 further upstream) be recalled or invalidated. Upstream DDS cache modules 20 use fs\_callback (often referred to as a slow ping) to register with the downstream DDS cache module 20 and provide a means for the delivery of asynchronous notifications.

When a slow ping is received, it is possible that multiple notifications are queued and waiting to be forwarded upstream. To handle such events expeditiously, the slow ping response can transmit multiple notifications.

The three preceding file system control operations provide the functionality required to ensure the integrity of file modifications, to implement cache consistency, and to quickly detect network partition events that compromise cache consistency.

The DDS protocol facilitates efficient transfers by allowing a single DDS\_LOAD or DDS\_FLUSH request to specify an array of file segments, each ranging in size up to 4 gigabytes, as targets of the request.

DDS\_LOAD and DDS\_FLUSH requests include flags that indicate whether the requesting DDS cache module 20 shares memory (DDS\_LOAD\_COMMON\_MEMORY) or disk (DDS\_LOAD\_COMMON\_DISK) with the downstream DDS cache module 20. Whenever data is being passed between DDS cache modules 20 with a common memory, pointers to the data are returned rather than the data itself.

A distributed consistency mechanism, an integral component of the DDS protocol and its implementation, enables a file's consistency control site (CCS—only exists when there's a concurrent write sharing condition present) to dynamically relocate itself as necessary ensuring that it is always positioned as far upstream from the DDS Server Terminator Site as possible but still able to monitor and coordinate all file access operations that occur while a concurrent write sharing condition is present.

The DDS protocol endeavors, with a minimum number of operations, to provide all the functions described above, and to thereby implement a superset of the functionality provided by all remote file access protocols. The protocol employs discriminated unions to virtualize the file object metadata that flows through and is cached within the DDS layer. Metadata is represented in its native format, and a

discriminator identifies the format whenever the metadata is referenced by a client intercept routine 42 in the course of responding to a file access request. This virtualization of metadata is the means that enables DDS to transparently service file access requests from unmodified client workstations regardless of the homogeneity/heterogeneity of the client with respect to the origin file server.

For example, in the process of responding to an NFS request, the NFS client intercept routine 42 must access the file's metadata. When the discriminator identifies the metadata format as NFS or UFS, an NFS client intercept routine ("CIR") can easily interpret the metadata and generate its response. However, when the metadata format is NTFS, an NFS CIR requires the services of an NTFS to UFS translation module in order to respond to the request.

#### DDS Domain Hierarchies

U.S. Pat. No. 6,847,968 B2 (the '968 patent") discloses the methods employed by DDS cache modules 20 to organize themselves into a hierarchy of domains. The '968 patent is hereby incorporated by reference as though fully set forth here.

FIG. 2 illustrates a DDS virtual file server for Inca Technology. As depicted, an inca domain 102 (with domain manager services provided by domain nodes inca and aztec) contains an eng domain 112, a sales domain 114, a corp domain 116 and a mrkt domain 118. The eng domain 112 (with domain manager services provided by domain nodes hoover and eureka) is non-atomic, which means it contains other domains (sub-domains). In this case the sub-domains are a bob domain 122, a joe domain 124, a pat domain 126, and a svrail domain 128. Three of these domains are atomic domains: the bob domain 122, joe domain 124 and pat domain 126 are all file servers, but svrail domain 128 is a non-atomic domain. The sales domain 114, also an atomic domain, consists of the resources being exported by a single file server. These resources include the exported file system 32A, the exported file system 32B and the exported file system 32C.

FIG. 3 illustrates a user's view of the inca domain 102. In this illustration the user is employing Microsoft's Explorer program to navigate through his computer's file space. Note that the DDS global file system has been mapped (connected to) the host computer's X: drive. As depicted, the X: drive contains two folders: inca\_mail and Internet. inca\_mail contains a private namespace and users must have the proper credentials to open that folder and view its contents. However, Internet is a public namespace and is open to all users.

Internet contains two top level domains: com and edu. The com directory contains cnn, hp, ibm, and inca. And, finally, the inca directory contains the inca domain tree depicted in FIG. 2. The cnn, hp, ibm, and inca directories are each the root of a company's domain tree. As FIG. 3 illustrates, when a domain tree root directory is opened (inca, in this case), the next level sub-domains (corp, eng, marketing, sales) appear.

Comparing FIG. 2's inca domain 102 with the inca folder in FIG. 3 clarifies the relationship between domains and folders: they are essentially the same thing. They are both resource containers that may recursively contain other resource containers and/or resource objects. A folder is a visual representation for a domain, and there may be other representations.

For example, the visual representation for a company's domain tree might be the corporate icon (with the well known filename logo.icon stored in the root directory of the domain tree).

FIG. 3 also depicts how DDS binds the shared resources of several companies (cnn, hp, ibm, and inca) into a single namespace enabling a user to seamlessly navigate across company boundaries.

The DDS namespace consists of two layers:

1. Filesystem Namespace—namespace defined by individual exported file systems. This layer is defined by the file systems (UFS, NTFS, EXT2FS, . . . ) containing the resource objects being exported through DDS.
2. Network Namespace—namespace consisting of DDS domain names. These names can usually be converted to an IP address using industry standard name resolution services such as domain name system ("DNS").

The Network Namespace ties together the disjunct namespaces of all the individual exported file systems to create a single namespace. DDS employs the existing network name resolution infrastructure to construct the Network Namespace. This results in the binding of exported file systems into the reference framework with which users and system administrators are already familiar.

FIG. 3 depicts the com directory as containing only four sub-directories. In reality, the com directory would contain the root directories of millions of company level domain trees. A single DDS virtual file server, encompassing the complete Internet namespace (gov, org, edu, mil, . . . ) and multiple private namespaces, may encompass hundreds, or even thousands, of petabytes.

This massive amount of data demands improved mechanisms for navigating through the global file system's namespace and for locating content of interest. Obviously, valuable content that cannot be located is actually valueless. DDS—A Step Beyond the Internet

The Internet as it exists today is an instance of a read only (mostly?) distributed file system on the same order of magnitude as what the DDS global file system will become. Today, Internet users routinely employ search engines to locate content of interest. These search engines appear to work quite well, but one should consider that users generally aren't aware of relevant content that a search fails to reveal.

The DDS global file system requires a search mechanism substantially faster and more efficient than the currently deployed Internet search engines. Recognize that DDS provides a file access service, complete with consistency guarantees. The Internet, by comparison, is an electronic distribution system for published content. Its content, once published, is unlikely to be modified. Furthermore, when an object is modified, a generous "grace period" is acceptable to allow the new content to migrate to distant access points (web proxy cache sites).

Even after most proxy cache sites have loaded the latest version of an object, it may be hours or even days before a web crawler fetches a copy to feed into an indexing engine. So, new content (and modifications to existing content) may not show up in search results for several days.

In contrast, the DDS global file system supports collaboration between individuals and groups. Whenever a document is created or modified, other users often need to be aware of the changes as quickly as possible. This gives rise to a requirement that DDS, to the maximum extent possible, index new and modified content in real time such that a search performed a few seconds after the creation of a new document will locate that document if it does, in fact, satisfy the criteria of the search.

DDS provides the functionally required to enable unmodified industry standard workstations to access remote files using their native CIFS or NFS implementations. DDS virtual file servers receive NFS or CIFS requests and service

them from cached data when possible and, when valid cached data is not present, DDS issues requests directed towards origin file servers to fetch the requested data.

Although the DDS protocol is highly streamlined and simplified in comparison with CIFS and NFS, it provides essentially the same capabilities. After all, the DDS protocol is designed to enable client access to files residing on very remote file servers. DDS implements the functionality provided by the file system APIs provided by Linux, Unix, Windows, and other major operating systems.

Using a DDS, NFS, CIFS, UFS, NTFS . . . API, an application establishes a connection to a content object through a series of operations:

1. Connect to a directory,
2. Enumerate the directory's contents,
3. Connect to the target object or to a directory believed to contain the target:
  - a. If connected to the target object: DONE.
  - b. If a sub-directory appears to contain the target object: GOTO Step 2.

Using the preceding method, a user can laboriously navigate throughout a file system and explore its content. However, discovering content by scanning directories becomes very inadequate when individual file systems encompass hundreds or thousands of petabytes of data. For such large file systems this method becomes unviable because users just don't live long enough. For large file systems, users (and processes) require more powerful methods for locating content, enabling them to quickly and efficiently establish connections to objects of possible interest.

For the Internet, the problem of searching the content of large files systems has already been addressed. Internet search engines accept user's queries and, in general, respond by providing the user with links to Internet objects that appear to contain something which meet a query's criteria. The user then uses the links to easily connect to objects of potential interest so their content may be perused and a final user determination made as to their relevance.

It is noteworthy that Internet search engines actually represent only the most recent instance of at least three successive generations of computer search engines which provide content searching:

#### 1. Dedicated Mainframe Systems

The early mainframe search engines (Dialog, Nexus, Lexus . . . ) indexed data residing on storage directly connected to the system hosting the search service. Users at terminals (both local to and remote from the mainframe) queried the system using a very structured Boolean syntax.

#### 2. Software Applications

Verity, Lexus, Nexus . . .

#### 3. Internet Search Engines

Lycos, Alta Vista, Magellan, Inktomi, Google . . .

Content based retrieval systems (search engines) are generally implemented as two distinct sets of applications:

1. Index generation applications, and
2. Retrieval applications.

Index generation is performed once on new (or modified) content, and the resultant new or updated indices are subsequently used by retrieval applications in responding to queries. Traditionally, content indexing systems, e.g., full text indexing, employ batch processing to index document collections. Dialog, one of the first mainframe based commercial full text retrieval systems, used nighttime hours for generating an inverted file which indexed its document collections, and during daytime hours provided online document search and retrieval services.

Presently, search engines continue to generate their inverted index structures in batch mode. Internet Web crawlers prowl sites, discover new content, and ship the new content back to index generation sites (which are usually sites also hosting search engines). New content, continuously flowing in from web crawlers, accumulate at the indexing sites. Eventually, the accumulated new content exceeds a threshold thereby causing it to be forwarded to an indexing engine. The indexing engine processes the new content, extracting and sorting index terms, and then merging the new terms into an inverted file. When invoked, the indexing engine processes all the accumulated new content in a single batch operation.

Although search engine technology has changed over the last thirty-five (35) years progressing from mainframe computer implementations to local area network ("LAN") implementations, and then from LAN implementations to wide area network ("WAN") implementations, the indexing component still retains its lineage: indexing is still performed as a batch mode process.

## DEFINITION OF TERMS

There appears to be no consensus about how terms associated with full text retrieval are used. Therefore, to avoid ambiguity some definitions for full text retrieval terms appear below:

**Document**—An object (file, record, document) within a collection associated with an accession number.

**Accession number**—A number, often assigned by the retrieval system's index generation software, which uniquely identifies a document within a collection.

**Linear file**—A collection of documents, concatenated together, often ordered by accession number.

**Linear file index**—An index into the linear file. Typically, each record in the linear file index consists of an accession number and a pointer to the associated document. In traditional full text retrieval systems such as Dialog, the pointer is a byte offset into the linear file. In Web based full text retrieval systems, the pointer may be a universal resource locator ("URL"). Records within this file are sorted by accession number.

**Index term**—A word or phrase extracted from a document during the parsing phase of the index generation process.

**Inverted file entry**—An index term followed by a pointer to a specific occurrence of the index term within a specific document.

**Inverted file record**—An index term followed by inverted file entries pointing to each occurrence of the index term within a document collection. The inverted file entries are ordered by <object ID, position within document>.

**Inverted file**—The complete set of inverted file records associated with a document collection.

Inverted file records may be alphabetically sorted by index term.

With Web based full text retrieval systems, the boundary between document collections has blurred. Distinguishing features of various document collections might be nothing more than that all the documents within a collection were indexed as a batch.

Individual documents might not be contained within a linear file; they might be geographically scattered about the Web and URLs within linear file index records provide links to these documents.

## 11

Inverted file index—An index into the inverted file. Typically, each record in the inverted file index consists of an index term and a pointer (byte offset) to the index term record in the inverted file.

Records within this file are alphabetically sorted by index term.

In some full text retrieval implementations, inverted file records are actually files. In which case, “inverted file”, as defined above, refers to the set of inverted files. And, in which case, there is no inverted file index since the containing file system provides the indexing required to locate an inverted file record.

Although indexing and retrieval isn’t the first thing that comes to mind when “file system” is mentioned, file systems do provide fairly complete name based (as opposed to content based) indexing capabilities. When an object is created, a new entry is created in a parent directory. The entry typically contains the object’s name and a link to the object’s attributes, which are stored within an inode (or equivalent). One of the inode’s attributes is an extent map specifying the device addresses (usually expressed as disk block numbers) where the object itself (the object’s data) is stored within the file system.

File systems generally use a hierarchical indexing structure that facilitates rapidly adding new entries and removing deleted entries. File system performance directly impacts overall system performance, so the speed at which entries can be created, deleted, and looked up has been a force that has molded all current file systems. In particular, name based indexing, which is fairly anemic when contrasted against content based indexing, is employed by all commonly deployed file systems.

File system developers have consistently and uniformly concluded that file system performance requirements exclude considering content based indexing. They’ve opted for speed over heavyweight indexing.

However, the file systems landscape has changed considerably over forty years. Consider the following:

1. Virtualized global file systems encompassing hundreds of petabytes are on the horizon if they are not already here, e.g. the World Wide Web.
2. These virtual file servers will consist of thousands of individual systems.
3. Individual systems may have very substantial physical resources:
  - a. Multiple CPUs, 4 Ghz and faster, 32 or 64 bit,
  - b. 2 to 64 gigabytes of main memory,
  - c. 1 to 1000 terabytes of disk memory,
  - d. Multiple GigE or 10 GigE network connections.

Incorporating a full text search capability into existing file system APIs as seamlessly as possible provides both users and processes with enhanced capabilities for locating, identifying and establishing connections based upon file content.

## SUMMARY OF THE INVENTION

An object of the present disclosure is to integrate a search capability into standard file system APIs such that existing programs, with little or no modifications, can transparently locate and discover content as an integral step in the connection establishment process.

Another object of the present disclosure is to provide a highly scalable distributed indexing capability.

Another object of the present disclosure is to provide a highly scalable distributed searching capability.

## 12

Another object of the present disclosure is to facilitate the extremely rapid indexing of new and modified objects so that these objects can be located based on their current content.

Another object of the present disclosure is to provide a method for generating a global scope object identifier which uniquely identifies an object contained within a global file system.

Another object of the present disclosure is to employ global scope object identifiers as accession numbers (see the preceding Definition of Terms) for objects indexed by a content retrieval system, allowing the inverted files generated in various sub-domains to be merged into a single composite inverted file.

Another object of the present disclosure is to distribute the functionality of a content retrieval system (such as full text retrieval) throughout the nodes of a global distributed file system such that generating (indexing) and manipulating (retrieving) of index terms is performed “close” to the content.

Another object of the present disclosure is to provide a means by which, for any given global scope object identifier, the object system (as in file system) containing the associated object can be quickly located even when the object system is frequently unmounted from one DDS object server and mounted on another such that it is essentially a wandering object system.

Another object of the present disclosure is to provide an extensible means by which the icons used in a graphical file management tool (such as Windows Explorer) to represent various domains in a global file system may be defined by an administrator for that domain.

Another object of the present disclosure is to provide an extensible, highly distributed, out-of-band signaling mechanism between processes concurrently accessing a common object.

Briefly, one aspect of the present disclosure is a method for incrementally indexing information contained in files within a distributed file system residing upon a virtual file server assembled by integrating a plurality of file servers. The indexing method includes:

- a. upon closing one of the files of the distributed file system after information contained in the file being closed has been changed:
  - i. parsing the information contained in the file; and
  - ii. extracting inverted index entries from the parsed information;
- b. sorting the inverted index entries; and
- c. merging the sorted inverted index entries into inverted file records of an inverted file that is associated with content of the distributed file system.

Another aspect of the present disclosure includes methods for ensuring that each inverted file entry uniquely specifies the source of the content identified by the inverted file entry, e.g. the particular file within a particular file system. The disclosed methods include generating a unique global object system ID for a file system included in a distributed file system residing on a virtual file server. One example of this is a 16 byte Universally Unique Identifier generated in accordance with RFC 4122. Alternatively, such a unique ID may be created by concatenating:

- a. the file system’s origin file server’s IP address; and
- b. the origin file server’s then current time. The global object system ID is stored within the file system (preferably within the superblock) and is the file system’s unique and permanent identifier. Lastly, the methods for ensuring that

13

each inverted file entry uniquely specifies the content's source include concatenating:

- a. the file system's global object system ID; and
- b. the object's object ID.

Systems and methods for extending the functionality of the consistency callback mechanisms employed by local file systems such as NTFS and distributed file systems such as DDS, NFS and CIFS so as to provide a shared memory foundation for efficiently broadcasting real-time content such as audio, video or high definition video from a source object to large numbers of viewers via the Internet are disclosed. Distributed applications such as video viewing client applications establish connections to a common distributed file system object, and then each application registers with the underlying distributed file system to receive notifications whenever the video source modifies the source object. The data required to update images maintained by viewing clients is included in notification messages. The distributed file system employs a network of proxy cache nodes. Proxy cache nodes receive notification messages (complete with image update data) and retransmit the messages towards the viewing clients using IP multicast techniques. In this manner, the distributed file system's consistency mechanism efficiently employs network resources to enable the real-time distribution of video content streams.

In one embodiment, a method for use in a distributed file system and for registering an application's request for notification of updates to a file comprises providing to the site at which the application is executing a file descriptor identifying the file for which the application requests notification of updates, and requesting that any notification of an update include the data that is updated in the file.

In another embodiment, a method of notifying sites of an update to a specified file comprises determining that there has been an update to the file, determining the sites at which the file is currently being accessed, and delivering a notification message to each of the determined sites indicating that the file has been updated and containing the updated data from the file.

In still another embodiment, a method of notifying an application in a distributed file system of an update to a specified file comprises registering at the site at which the application is executing a descriptor identifying the file for which the application requests update notification; receiving at the site a notification message that an update to the file has occurred and the newly updated data in the file; and delivering the notification to the application.

In still another embodiment, a method of communicating updates to a file located at a first site in a distributed file system to applications running at different sites in a distributed file system comprises: for each of a plurality of applications, providing to the site at which the application is running a file descriptor identifying the file for which the application requests notification of updates; for each file for which a file descriptor is provided, storing the request for update notification as extended attributes of the identified file that are handled in the same manner as the file's regular attributes; receiving an update to the file; determining the sites at which the file is currently being accessed; delivering a notification message to each of the determined sites, the notification message indicating that an update has occurred and containing the newly updated data in the file; and, at each site delivering a notification and the newly updated data to the application that provided the file descriptor.

In yet another embodiment, a method of broadcasting source content located on a first site in a distributed file system to multiple applications running at a plurality of

14

other sites, comprises: for each of the plurality of applications, establishing through the site at which the application is running a connection to the first site or to an intermediate site which is closer to the first site; for each intermediate site, establishing a connection to the first site or to another intermediate site until a connection to the source object located on the first site is established for each of the plurality of applications; returning to each of the applications an object identifier identifying the source object; buffering a copy of the source object at each site at which an application is running and at each intermediate site; for each of the plurality of applications, requesting an update notification whenever the source object is modified; determining when new data has been written to the source object and in such instances providing an update notification to each of the plurality of applications by sending the update notification to each of the next level sites through which the plurality of applications are connected to the source object, the update notification including an indication of the new data written to the source object; and updating each buffered copy of the source object with the new data.

These and other features, objects and advantages will be understood or apparent to those of ordinary skill in the art from the following detailed description of the preferred embodiment as illustrated in the various drawing figures.

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a functional type block diagram depicting a DDS cache module;

FIG. 2 is a tree diagram depicting several hierarchical domain trees; and

FIG. 3 is a user view display depicting file space for the several hierarchical domain trees illustrated in FIG. 2.

FIG. 4 is an illustration of an event descriptor control block.

FIG. 5 is an illustration of a notification descriptor message.

#### DETAILED DESCRIPTION

Embedding a full text search engine into a distributed file system to automatically index the file system's content requires that the search engine:

1. Integrate seamlessly into virtual file server frameworks.
2. Be capable of making new content immediately searchable, i.e. obviate a need to perform a separate batch process to index the file system's content.
3. Be highly scalable—during both index generation and retrieval operations.

Consequently, integrating a content based retrieval system into a distributed file system breaks down into separate tasks of integrating an index generation capability and integrating a content retrieval capability into the distributed file system.

##### Integrating Index Generation

New content becomes instantly searchable when content indexing is integrated directly into the main code path of the software routines implementing a distributed file system. However, the sheer volume of information stored in large distributed file systems demands more powerful indexing capabilities. Fortunately, the resources are now available to provide very robust indexing at the demanding speeds required by file systems. However, content based indexing requires such a substantial amount of additional processing that speed and efficiency must be the prime directives shaping its inclusion in a distributed file system. In particular, speed and efficiency dictates that inverted file index

generation should be performed at a site as close to the object as possible, preferably at the same site as where the object is stored.

The preferred method by which new content becomes immediately searchable is:

1. Shortly after the last byte of data has been written to a newly created (or modified) object, the client process (creating the object) invokes the file system's API close routine. In traditional file systems (hierarchical name based indexing), only a few attribute updates (file size, modification time) are required to "finalize" some of the attributes associated with the new object. These updates are performed as usual.

2. Then, the new object is immediately parsed. Potentially, up to every word (excepting "junk" words like and, or, the, . . . ) is extracted and converted into an inverted index entry.

The parsing phase generates a number of index term buckets. Each bucket is labeled with an index term extracted from the object, and the bucket's contents are a variable number of inverted index entries. For example, if the word "computer" occurred three times in the object, there would be three entries in a bucket labeled "computer." Each entry in the index term buckets identifies where the particular word on the bucket label, e.g., "computer," occurs within the object.

The format of an inverted index entry varies between different full text indexing systems. Described most generally, an inverted index entry has the form <object ID, offset within the object> which typically takes the form <object ID, field, sentence, word>. For example, if "computer" occurred as the seventh word in the second sentence of the abstract field (where abstract field equals 121), then its entry would be <236741.44127366, 121, 2, 7>. (Obviously, the new object has been assigned an identifier of "236741.44127366".)

3. After the parsing phase completes, the inverted index entries in each bucket are sorted.

4. After sorting, the inverted index entries contained in the index term buckets are merged into the inverted file. For example, the three sorted entries in the "computer" bucket are merged into the "computer" inverted file record, which contains the sorted set of inverted index entries that specifically identify each occurrence of "computer" throughout the complete set of objects that had been previously indexed.

5. After the merge completes, the object is fully indexed and can be located using content based searches.

6. The close operation may return immediately after initiating a new thread to perform the full text indexing of the new object, or it may return only after the full text indexing has completed. In the latter case, the application that just created (or modified) the object is assured that the object is completely searchable from the moment the close function returns.

The preceding process updates the inverted file incrementally, one object at a time. This is a departure from the batch mode updates (where thousands of objects are typically processed at a time) that are usually employed by today's full text retrieval systems. While updating the inverted file incrementally may not be new, its integration into a file system's close operation departs significantly from known file system practice.

#### Index Generation Scalability Issues

The process of generating an index for a collection of objects requires that every object be scanned and parsed

(index terms extracted). Following the parsing phase, the inverted index entries are sorted and merged into the inverted file. This process is both cpu intensive and i/o intensive.

As enumerated above, the full text indexing of a new object may require two, three, or even four orders of magnitude more processing than the traditional directory hierarchy style indexing. This is a marked increase in the processing required during a file system's close operation. Therefore, full text indexing must be invoked judiciously. There may be many instances where it's known beforehand that full text indexing isn't warranted for certain types of objects, e.g. files containing pictures or images.

The invocation of full text indexing during a file system's close operation is optional and is controlled by either a flag bit parameter (DDS\_GENERATE\_INDEX) of the close function call and/or the definition of a new file system API: close\_and\_index.

A flag bit parameter (DDS\_SYNCHRONOUS\_INDEXING) to the close or close\_and\_index functions determines whether the full text indexing is completed before the file system's close operation returns to its caller.

#### Differential Inverted File Updates

Another way to increase index generation scalability is to reduce the amount of work required to update the indices for a file following its modification.

Consider a book on "computer networks" being written by several authors. Assume each author is writing a chapter (in a separate file) and all authors are interested in changes to any chapter being visible to all of them.

After a particular co-author modifies Chapter 7 and the close\_and\_index routine is called, the following sequence of operations occurs:

1. The chapter is parsed and chap7.drev11.if, an inverted file for the eleventh revision of Chapter 7, is generated.
2. chap7.drev11.if is differenced against chap7.drev10.if to generate the difference file chap7.ddiff.10-11. This file indicates which terms were added, deleted, or moved during the last editing session.
3. A book.if composite inverted file, containing the merged full text indices for the complete book, is updated by applying all differences listed in chap7.ddiff.10-11.

Without a differential update capability, the sequence would be:

1. The chapter is parsed and chap7.drev11.if, an inverted file for the eleventh revision of Chapter 7, is generated.
2. The book.if composite inverted file, containing the merged full text indices for the complete book, is created by merging the latest versions of each chapX.revX.if.

The advantage of differential updates becomes apparent when the composite document collection consists of millions of documents. The task of selectively updating book.if in this case is substantially less than the effort required to merge millions of xxxx.if files.

#### Integrating a Content Retrieval Capability

A preferred method for seamlessly integrating a full text search capability into the DDS virtual file server framework is to incorporate a content retrieval capability into DDS and to extend DDS's verb set (distributed file system API) to make this new capability accessible to DDS clients.

In addition to the standard file system verbs (open, close, read, write, link, rename, . . . ), a new verb is added: search. DDS clients may invoke a full text retrieval mechanism, the preferred content retrieval capability, by calling upon the new search file system API routine.

The response to a successful search request is a set of inverted index entries. As described previously, an inverted index entry generally has the form <object ID, offset within the object> which typically takes the form <object ID, field, sentence, word>. The object ID of an inverted index entry identifies an object, and the other information within an inverted index entry points to a location within the object where the index term associated with the entry occurred.

An object ID (accession number) may have either local or global scope:

1. Local scope object IDs uniquely identify objects within the domains that contain the objects. A local scope object ID is not valid and cannot be interpreted outside of its domain. However, a local scope object ID may be used by foreign client processes (operating outside of the local scope) as an opaque handle for the object it identifies. A foreign client process may present an opaque handle to a process executing within the local scope and request that the local process establish a connection to the object.

However, when a global domain contains millions of sub-domains, it is not always readily apparent which sub-domain is the progenitor of a particular local scope object ID.

2. Global scope object IDs uniquely identify objects throughout a global domain, and are universally valid. A very important property of universal validity is that all of the inverted files (created mostly in sub-domain leaf nodes) of the global domain comprise a single distributed inverted file. This has the following implications:
  - a. The actual execution of a global search may be performed in the individual sub-domains and the results from each sub-domain may be merged into a single response.
  - b. The inverted files generated in sub-domains may be loaded by the parent domain and merged into a single composite inverted file. After the merge, searches executed in the parent domain using this single composite inverted file will yield all relevant content contained in the parent domain and all of its sub-domains.

#### Global Scope Object ID

Indexing content within objects of a distributed file system requires that the object ID included in each inverted file entry uniquely specify the source of the content identified by the inverted file entry, e.g. the particular file within a particular file system. Ensuring the existence of unique object IDs globally throughout an entire distributed file system requires that the distributed file system, at least, maintain a registry of unique identifiers, e.g. unique file system identifiers.

Basically two approaches exist for registering unique identifiers:

1. a proposed identifier is generated and submitted to the registry which either accepts or rejects the identifier.
2. the registry generates and issues a unique identifier in response to a request therefore.

Currently, file systems for Unix, Solaris and Linux are capable of generating object identifiers that are essentially unique within a particular file system. A preferred method for generating global scope object IDs is to prepend to this particular type of file system object ID a globally unique object system ID. Accordingly, a global scope object ID

which uniquely identifies an object throughout a global domain preferably has the following structure:

```
global scope object ID ::=
  <global object system ID><object ID>
Global Object System ID
```

One type of global object system ID is a sixteen byte UUID (Universally Unique Identifier) generated in accordance with RFC 4122 when a file system (the most prevalent type of object system) is created by, for example, either a newfs or mkfs Unix or Solaris command. However, any unique id may be used, for example, a twelve byte number in which the first four bytes of the global object system ID are the origin file server's IP address, and the next eight bytes are a high resolution time stamp (microsecond or better) provided by the origin file server which reflects the current time. As mentioned above, the global object system ID may be generated on the host computer creating the new file system or the host computer may request the registry to generate, register and return a global object system ID.

The first step in creating a new file system is to generate a UUID in accordance with RFC 4122, or other number as described above, and use the resulting number as a candidate for the global object system ID for the file system about to be created. Next, an attempt is made to register the candidate global object system ID with a DDS provider locator service (described below). If the global object system ID is already in use, the locator service responds with an error.

If the DDS provider locator service refuses to register a global object system ID because it is not unique, the origin file server once again generates either a UUID in accordance with RFC 4122 or other number as described above. Then this new and different global object system ID is submitted for registration with the DDS provider locator service. This iterative process repeats until the origin file server successfully registers a global object system ID with the DDS provider locator service.

Having successfully registered a global object system ID with the DDS provider locator service, the origin file server proceeds with creating the new file system. After the file system is created, the registered global object system ID is entered into the file system superblock, thereby assigning a unique and permanent identifier to the file system.

It is readily apparent to those skilled in the relevant art that there exist many other techniques which may be employed for generating and registering a unique global object system ID.

Object ID An object ID is a number created by a provider for the purpose of uniquely identifying an object. An object ID may be temporary or it may be permanent, and some providers generate both types.

For example, on a Solaris file server a local process might open the same file on two different occasions and receive two different file descriptors: "11" the first time and "7" the second time. (A file descriptor is one type of object ID.)

However, a remote NFS client accessing the same file would receive a file handle that uniquely identified the file forever. An NFS client may present a file handle that it received ten years ago and hasn't used since back to the file server and that server must either establish a connection with the original file or respond with an error indicating that the file handle is no longer valid. (A file handle is another type of object ID.)

A method commonly used by Unix based NFS file servers to create a permanent file id is to concatenate two 32 bit numbers, the inode number and the inode generation number, to create a 64 bit file id. Since each time an inode is assigned to a new file its generation number is incremented, an inode would have to be re-used over 4 billion times before a file id of this type could repeat. These 64 bit file ids are essentially good forever.

19

## DDS Provider Locator Service

The provider locator service is a network directory service. The locator service tracks the current location of a file system within a geographically distributed virtual file server as the file system is repeatedly unmounted from one file server and then mounted on another server.

Whenever a new file system is mounted or unmounted, a message is sent to the DDS provider locator service informing it that “file\_system\_X” (identified by the global object system ID) is being mounted (unmounted) on “host\_X” (a fully qualified domain name that can be resolved to an IP address).

Combining a method to create global object system IDs (such as the one described above) with a method of determining which file server currently has a specified file (object) system mounted effectively decouples an object ID from the server that created it.

Windows, Linux and Unix file servers generate local scope object IDs which can only be interpreted by the servers that generated them. The method described here enables file servers to interpret the object IDs for files contained within all currently mounted file systems regardless of which file server originally generated the object ID.

Consider a DDS client with a ten year old file handle. Suppose that the file associated with the old file handle resided within a file system on a disk drive that had recently been shipped to a distant location. The disk is subsequently installed on a DDS configured file server and that server mounts the file system. Then, when the DDS client attempts to read the file, the “old server” responds with an error indicating “no such file system”. This causes the client system to query the locator service, requesting the name of the server currently providing access to “file\_system\_X”. The client then re-directs the read request towards the “new server”.

The combination of global object system IDs and the DDS provider locator service enables object system migration without client disruption. This capability facilitates and greatly simplifies implementing failover, load balancing, archival, and disaster recovery type products.

In one embodiment, when a file system is created:

1. A name (for example eng1673, up to 64 chars) is stored in the superblock.
2. A global object system ID (RFC 4122 type UUID, 16 bytes) is generated and stored in the superblock (for example: 550e8400-e29b-41d4-a716-446655440000).

When the file system eng1673 is mounted on a file server, for example prima-donna:

1. The superblock is read and the file system’s name and global object system ID are retrieved.
2. The provider locator service is queried (using LDAP or equiv) to ensure that the file system is not currently mounted. If the file system is already mounted, the mount operation terminates and reports this error condition. (This error may be the result of the file system being cloned for backup purposes and then an attempt is made to mount the backup copy of the file system while the primary is mounted.)
3. The mount point directory is read. This directory, which is specified in the mount request, preferably has the same name as the file system that is being mounted. For example, the mount point directory for eng1673 would preferably be a pathname with the last component being ‘eng1673’ (such as /remote/eng1673).

20

A. If a file named 550e8400-e29b-41d4-a716-446655440000 exists within the /remote/eng1673 directory, the file system has been mounted here before.

B. If a file named 550e8400-e29b-41d4-a716-446655440000 does not exist, the file system has not been mounted here before. The file ‘550e8400-e29b-41d4-a716-446655440000’ is created at this time, and then a second name ‘eng1673’ is created for the file (using a Unix hard link). The file contents may include, for example, the file system’s:

- a. Name, i.e., eng1673,
- b. global object system ID, i.e., 550e8400-e29b-41d4-a716-446655440000,
- c. time\_of\_last\_mount, i.e., current time,
- d. time\_of\_last\_unmount, i.e., -1, and
- e. forwarding action, i.e., REFERRAL.

4. The mount operation is performed, thereby “covering” the file named 550e8400-e29b-41d4-a716-446655440000 (also named eng1673) that was possibly just created.

5. The provider locator service entry for prima-donna is updated using LDAP or the equivalent:

- A. the eng1673 value is added to the fs\_name attribute, and
- B. the 550e8400-e29b-41d4-a716-446655440000 value is added to the fs\_uuid attribute.

In one method, to unmount the file system eng1673 on file server prima-donna:

1. The network directory service entry for prima-donna is updated:
  - A. the eng1673 value is deleted from the fs\_name attribute, and
  - B. the 550e8400-e29b-41d4-a716-446655440000 value is deleted from the fs\_uuid attribute.
2. The file system is unmounted using normal procedures for this type of file server. (Unmounting the file system “uncovers” the file eng1673.)
3. The file 550e8400-e29b-41d4-a716-446655440000 is updated with the time\_of\_last\_unmount, i.e., current time.

In a corresponding embodiment, when a file system migrates:

1. The file system is unmounted using the procedure outlined above.
2. The file system is then mounted on a new server (which may be geographically remote) using the procedure outlined above. During the mount operation the file system’s new location is registered with the provider locator service.

Later, when a remote client attempts to access a file within the migrated file system using a pre-migration pathname:

1. The DDS service running on prima-donna receives a DDS\_CONNECT request targeting a file or directory within the root directory of eng1673.
2. The migration of file system ‘eng1673’ is detected when:
  - A. the DDS service fails to establish a connection to the targeted file or directory within eng1673 because it does not exist; and
  - B. the file 550e8400-e29b-41d4-a716-446655440000 is discovered within the /remote/eng1673 mount point directory.
3. The file 550e8400-e29b-41d4-a716-446655440000 (now uncovered) is read and the name is verified to be ‘eng1673’ and then the forwarding action and global object system ID values are retrieved.



## 21

4. Depending on the forwarding action, several events may occur:

- A. REPORT\_ERROR:

The system may return an error message, such as ERROR\_FILESYSTEM\_NOT\_MOUNTED.

- B. REFERRAL:

- a. The provider locator service may be queried and the IP address (or name) of the server now hosting file system 550e8400-e29b-41d4-a716-446655440000 is retrieved.

- b. If the host server's name was retrieved, DNS (or equivalent) may be called to convert the name into an IP address.

- c. A referral response is dispatched back to the client system and the client transparently redirects the file access request towards to file server currently exporting 550e8400-e29b-41d4-a716-446655440000.

- C. TUNNEL:

- a. The provider locator service may be queried and the IP address (or name) of the server now hosting file system 550e8400-e29b-41d4-a716-446655440000 is retrieved.

- b. If the host server's name was retrieved, DNS (or equivalent) is called to convert the name into an IP address.

- c. The request is forwarded to IP address.

- d. Any received response is passed back to the client.

#### Searching Using an Existing File System's API

The preferred method incorporates a search capability in a way that enables unmodified application programs to employ the new capability. The following example illustrates how this is accomplished:

1. On a hal workstation 132, a Windows 2000 workstation, an application program executes the system call: CreateDirectory(lpPathName, lpSecurityAttributes); where lpPathName is a pointer to the string: "X:\Internet\com\inca\!\_Search\_Results\_!(computer OR PDA) AND network", and lpSecurityAttributes is a pointer to a SECURITY\_ATTRIBUTES structure.

Both the PathName and the SecurityAttributes parameters conform to the Windows 2000 conventions.

2. The hal workstation 132's X: drive is mapped to a DDS portal (a DDS node on the same LAN segment). Therefore, the Windows 2000 Redirector on the hal workstation 132 generates a CIFS CREATE\_DIRECTORY request and dispatches it to the DDS portal. The request contains the lpPathName string as the CIFS DirectoryName request parameter.

3. The request arrives at the DDS portal and is processed by DDS's create directory CIFS CIR. Request processing initially proceeds as normal:

- a. The Internet directory is opened.
- b. The com directory is opened.
- c. The inca directory is opened.

4. However, when the !\_Search\_Results\_! pathname component is encountered, the character string is recognized as a special directory name where the roots of parallel namespaces created by searches of the inca domain 102 are anchored (see Navigable Namespaces for Search Results—Parallel Namespaces below).

5. An attempt is made to open the !\_Search\_Results\_! directory, which will exist if prior searches of the inca domain 102 have been performed.

6. The !\_Search\_Results\_! directory is created if it does not already exist.

## 22

- 7.d The pathname component following "!\_Search\_Results\_!" is interpreted as a search specification string. The integrated content search mechanism is invoked by an inca domain 102 manager (either inca or aztec) at this point by DDS's CIFS client intercept routine to search the \Internet\com\inca domain according to the search specification.

8. A directory with the name "(computer OR PDA) AND network" is created within the !\_Search\_Results\_! directory. This directory, the root of the parallel namespace created by the search, contains links to the content within the inca domain 102 that satisfy the search criteria.

9. When a search is performed and no content matches the search specification, an error (NON\_EXISTENT\_DIRECTORY) may be returned or the search root directory may be created. However, it will be empty; indicating that there were no matches for the query.

#### Extending the Current File System API

Another method of extending standard file system APIs is to define a new function: the search function. The search function's first parameter is a search specification text string, and the remaining parameters are the same as those of the CreateDirectory function. One of those remaining parameters is the pathname of the domain to be searched.

When a search produces results, a directory is created and it is immediately populated with links to the results. The caller opens the new directory and follows the links within to establish connections to objects that satisfied the search criteria.

#### Content Retrieval Scalability

An integrated content search capability gives file system clients a powerful new tool for locating content. However, placing this tool within easy reach of users (and there are many users!), creates some major scalability issues.

Given a new search verb, it's imperative that DDS respond to search requests quickly and efficiently. Not only do DDS virtual file servers encompass huge pools of information, but they also support millions of users concurrently. Since thousands of search requests may be in progress simultaneously, searches must be performed relatively quickly and must not excessively burden the network. However, responding to a search request generally places a far heavier burden on a virtual file server's infrastructure than any of the other file access requests a client might submit.

The magnitude of the content retrieval scalability issues completely dwarfs the issues associated with real-time full text indexing.

Consider a single user submitting a search request specifying /dds/Internet/com as the root of the subtree to be searched:

- a. Since this is a request to search all content contained by all of the organizational domain trees plugged into the /dds/Internet/com domain, it must be forwarded to each of the millions of domain trees populating /dds/Internet/com.
- b. Many organizational domain trees (such as /dds/Internet/com/boeing) are, in themselves, enormous geographically distributed domains consisting of thousands of industry standard file servers.
- c. There are likely to be tens of thousands of such search requests in process simultaneously.

Clearly, effectively addressing the content retrieval scalability issues is an enormous task. Accordingly, a variety of methods, each of which focuses on some aspect of retrieval scalability, are employed to implement fast, responsive and

highly scalable retrieval capabilities. Each of these various methods can be categorized by its basic approach to the problem.

The methods, grouped by category, are:

1. Methods to increase search engine scalability:
  - a. Hierarchically distributed search engines
  - b. Inverted file interoperability
  - c. Inverted file caching
  - d. Managed consistency:
    - i. Static domains
    - ii. Energetic domains
    - iii. Delayed consistency
    - iv. Selective consistency
2. Methods to limit the distribution of search requests:
  - a. Public namespace windows
  - b. Domain level access control
  - c. Search request routing
3. Methods to support public search engines:
  - a. Inverted file forwarding
  - b. Inverted file caching
4. Methods to create navigable namespaces for search results:
  - a. Attribute extensions
  - b. Parallel namespace resolution
  - c. Namespace Discrimination
  - d. Namespace Coherency

The following sections and sub-sections follow the immediately preceding categorization, and provide methods within each category.

#### Search Engine Scalability

DDS global file systems are organized as domain hierarchies. If a single monolithic search engine were used to implement a global file system's search function, that search engine would be continuously burdened with the processing of every search request generated by all users, and the number of users at any one time might number in the tens of millions. Obviously, this monolithic approach lacks scalability.

#### Hierarchically Distributed Search Engines

An alternative to a monolithic solution is hierarchically distributed search engines. This approach delegates the responsibility (burden) of performing searches down into the sub-domains. And, each sub-domain may, instead of actually executing any search routines, rely on its sub-domains to provide the search capability for each of its sub-domain's respective resources. In this way delegating search request processing to sub-domains becomes recursive. Atomic domains, leaf nodes which do not contain sub-domains, must either service the request or respond with an error code indicating that this domain does not provide a retrieval capability.

When a domain receives a search request, the domain may:

- a. process the request directly, or
- b. forward the request on to each of its sub-domains, or
- c. forward the request on to some of its sub-domains while also processing the request directly for itself and for some of its sub-domains.

#### Inverted File Interoperability

When a domain generates a response to a search request, it merges the responses received from its sub-domains (if any) with its own response (if any).

The complete flexibility to execute searches at various levels of the hierarchy and to merge the results derives from the exclusive use of global scope object IDs to identify all objects within the namespace.

#### Inverted File Caching

Inverted files are usually generated in atomic domains (leaf nodes), where most of the content within a global domain resides. Generating indexes in the leaf nodes:

- a. Keeps a substantial amount of file access traffic from flowing across network links.
- b. Distributes the indexing burden as widely as possible.

When the eng domain **112** in FIG. **2** receives a search request, the search may be directly executed by the eng domain **112**, as opposed to forwarding the request on to the joe domain **124**, pat domain **126**, bob domain **122** and svrail domain **128**. The inverted file(s) referenced during the execution of the search request may be local (having been previously pulled from the sub-domains and merged), or they may be remote. In the latter case, DDS's caching mechanism automatically loads images of those portions of each sub-domain's inverted file that are relevant to the searches being executed in the parent domain. However, in this case, the search must be individually executed for each sub-domain and the results must be merged.

Consider the inca domain tree depicted in FIG. **2**. Let's step through the processing of a search request. Referring back to FIG. **3**, a user decides to search the inca domain **102** for all resources indexed by both "computers" and "networking". The user right clicks on the inca folder and selects a "Search . . ." option which appears in a pop-up menu.

A Search Dialog Box pops-up and the user enters "computers AND networking" and initiates the search, resulting in the following system call being made:

```
search("computers AND networking", "/dds/Internet/
com/inca", lpSecurityAttributes);
```

The search system call eventually results in a search request arriving at one of the inca domain portals. (A DDS domain portal is an access point for the domain's resources. It is an IP address where the DDS service provides access to the domain's resources.) This request is handled by a search engine integrated into DDS, which may:

1. Process the request directly at this site, referencing a single inverted file that "covers" all objects in this domain and all sub-domains. The single inverted file:
  - a. may have been generated at this site by indexing the content of each sub-domain at this site, or
  - b. may have been generated with the substantial cooperation of the sub-domains, each sub-domain indexing its own data and forwarding the resultant inverted file to this site where it's merged into a single composite inverted file.
2. Process the request directly at this site, referencing remote inverted files residing in each of the sub-domains. This mode, which employs DDS's caching mechanism to make the relevant parts of remote inverted files appear to be "here", must execute the search individually against each sub-domain's inverted file. The results of the individual searches are then merged into a single response.
3. Forward the request on to each sub-domain. The responses received back from the sub-domains are merged into a single response.

A \dds\Internet\com\inca!\\_Search\_Results\\_!computers AND networking\ directory is created (if it does not already exist) for storing links to the objects identified by the search, and a file descriptor for the directory is returned to the caller. Using the returned file descriptor, the caller may read the directory's contents and thereby begin the process of exploring the parallel namespace (created by the search engine, see Navigable Namespaces for Search Results—Parallel Namespaces below) populated with the results of the query.

"!\_Search\_Results\_!" is the root of a parallel namespace dynamically created by the integrated search engine in response to queries. Its purpose is to keep all search results in a separate namespace, with only the portal to that namespace visible to users. When the user lists the contents of the \dds\Internet\com\inca directory, his search results (and those of other users) are not displayed; only the !\_Search\_Results\_! directory is shown. However, the user can easily navigate into the !\_Search\_Results\_! directory to view the results of his prior searches and possibly the results of prior searches performed by other users (if allowed by the access permission attributes of the prior searches).

A parallel namespace contains a subset of the objects contained in a primary namespace; the subset composed of objects that matched a content retrieval query. For example, "computers AND networking" is the parallel namespace root directory created by the content retrieval query: "computers AND networking". The "computers AND networking" directory is contained within "\dds\Internet\com\inca\!\_Search\_Results\_!", which is the root directory containing all parallel namespaces created by searches of the inca domain 102.

The !\_Search\_Results\_! directory also serves an administrative purpose. Search results are usually automatically deleted on a least recently used basis whenever DDS decides it's time to recover disk space. Keeping all search results under a single directory facilitates recovering disk space.

#### Managed Consistency

When inverted file caching is employed, the consistency maintained between a source inverted file and its projections into parent domains may be subject to the mechanisms described in this section.

Searches may be executed by referencing remote inverted files contained within sub-domains. DDS's caching mechanism is very effective at making static remote inverted files (data that rarely changes) appear to be local. However, when remote inverted files are modified frequently, the network traffic required to maintain cache consistency (completely current indices) becomes substantial. The following sections present strategies for executing searches when the content is changing at various rates.

#### Static Domains—Execute Searches in Parent Domains

Static domains contain documents that rarely, if ever, change. For example, the published documents domain of a large corporation might contain employee manuals, company procedures, annual reports, product specifications and other documents that have been the subject of a formal review process. These documents are fairly static; they don't tend to change very often. The inverted files generated from these documents also tend to be static and are excellent candidates for caching in parent domains. The implication here is: A static domain may be efficiently searched by an engine executing on a parent domain host node.

Referring to FIG. 2, the pat domain 126 and joe domain 124 (sub-domains of the eng domain 112) are static domains. A search executed on hoover (a domain manager node of the eng domain 112) could reference remote inverted files generated by, and contained within, the pat domain 126 and joe domain 124. Those segments of the remote inverted files accessed by the search process executing on hoover would be cached locally.

Assuming the availability of cache space, hoover's cache would eventually be populated with whatever indices are required to support the type of queries handled by this node. These indices might be a small subset of the complete set of remote inverted files.

#### Energetic Domains—Execute Searches within the Domain

Energetic domains represent the other end of the spectrum. When a domain's content is constantly changing, the associated inverted file is subject to continuous modifications. These modifications render inverted file caching within parent domains ineffective. So, for domains with volatile content, a good rule is: An energetic domain is most efficiently searched by an engine executing on a domain host node.

Referring to FIG. 2, the bob domain 122 and svrail domain 128 are both energetic sub-domains of the eng domain 112. A search request received by hoover (an eng domain 112 manager) may be forwarded to the bob domain 122 and svrail domain 128 and then hoover may perform the search for the pat domain 126 and joe domain 124. After all four searches have completed, hoover merges the results into a single response.

#### Delayed Consistency

Energetic domains may be efficiently searched by remote processes by relaxing the consistency constraints on cached image projections of inverted files. Instead of ensuring that a read request always returns the most current data, the consistency mechanism can be set such that a read request returns data that was current as of an hour ago.

Suppose "computer" is one of the search terms comprising a query. Employing relaxed consistency, when the search engine references the "computer" indices cached at this site, it checks an associated timestamp indicating when the "computer" indices data was last refreshed. New indices are loaded only if the cached image is out of date by more than a specified threshold (e.g. one hour).

Of course, many variations of delayed consistency can be conjured up to capitalize on the fact that, in most instances, a search that discovers all relevant content based on indices generated relatively recently is sufficient.

#### Selective Consistency

However, there are situations where collaborative efforts between numbers of individuals (or processes) demand that content modifications be instantly visible to all members. The cost associated with maintaining this level of visibility is substantial, so a mechanism is provided to selectively deliver the higher levels of indexing consistency.

A distributed search routing switch determines where a search is actually executed within a global file system hierarchy. Referring once again to FIG. 2, let's step through the processing of a search request:

1. A search request is received by hoover, one of two DDS nodes hosting the eng domain 112. ("Hosting the eng domain 112" means "providing the domain manager services for the eng domain 112".) The domain contains:
  - a. The joe domain 124 and pat domain 126—static domains. Their cached indices are completely current.
  - b. The bob domain 122, and svrail domain 128—energetic domains. Their content is always changing. However, recently modified objects usually comprise only a small fraction of total content. So, with delayed consistency set to one hour, 99.5 percent of the cached images of the bob domain 122's and svrail domain 128's indices are likely to satisfy that level of consistency. (Only modifications made during the last hour will not be reflected in the cached images of the bob domain 122's and svrail domain 128's indices.)

27

2. The request credentials identify the user (or process) as a principal collaborating on a project contained within the bob domain **122**. With respect to the bob domain **122**'s content, this user wants recent modifications to be as visible as possible.
3. The eng domain **112** manager hoover initiates four search threads, each thread targeting a different sub-domain (the pat domain **126**, joe domain **124**, bob domain **122**, and svrail domain **128**).
4. Each thread first calls the distributed search routing switch, a software routine that determines whether:
  - a. the search should be executed at this node, or
  - b. the search should be forwarded on to the sub-domain.
5. The distributed search routing switch determines that the threads targeting the two static domains (the pat domain **126**, joe domain **124**) and the thread targeting the energetic domain in which the user's not particularly interested, i.e. the svrail domain **128**, should each invoke the search engine at this node (hoover). These three threads will execute the search by referencing the remote (and possibly cached) inverted file indices of their respective target sub-domains. (When the inverted file indices are already cached, they may be slightly stale because consistency constraints may have been relaxed in the interest of reducing network traffic.)
6. The distributed search routing switch also determines that the search of the bob domain **122** should be executed directly on the bob domain **122**. Completely current indices (bob's inverted files) are to be used for this search because this user wants recent modifications of the project to be instantly reflected in the search results. So, the fourth thread forwards the search query on to the bob domain **122** and waits for the response.
7. Each thread generates a response if any content within its targeted domain matches the query. Each response is a derived set of inverted index entries; i.e., each set is the result of a Boolean set operation performed on input sets of inverted index entries. Each input set may have been created by the index generation routines or it may be a derived set from previous Boolean set operations. The global scope object ID (accession number) contained within each set entry is a link to an object matching the query. Thus, each response is a set of links to the objects within a sub-domain that were identified by the search thread.
8. The response generated by the eng domain **112**'s manager hoover is patterned after a DDS\_CONNECT type response. The response includes the file handle and attributes for an object contained within the \dds\Internet\com\inca\!\_Search\_Results\_! directory. The object may be either a directory or a file (as indicated by the attributes). The object may have been created in response to this search request or it may have been created by a previous request. New data fields added to the attributes may indicate the number of hits and/or a relevance weighting.

In this case, the file handle for a directory is returned by hoover. When the calling process reads the directory, it will discover three files (named joe, pat, and svrail) and a directory (bob). Each file contains links to sub-domain objects matching the query. The bob domain **122**'s directory contains a remote file named bob. This file is just like the others except it actually resides on bob domain **122** whereas the other three files reside on hoover.

28

For efficiency reasons, the eng domain **112**'s manager hoover may optionally merge all four files into a single file and then return the file handle for that file instead of returning a directory file handle.

#### 5 Distribution of Search Requests

A single unrestricted search (the only kind described so far) targeting the com domain would generate a global wave of network traffic. The search request, delivered to the millions of domains plugged into dot\_com, would recursively propagate from domains to sub-domains until every node within the com domain had participated in the search.

Obviously, the amount of network activity required to satisfy a single user request is excessive and doesn't scale beyond a relatively limited number of users.

Therefore, several methods may be employed to limit the distribution of search requests targeting public namespaces (such as com, edu, gov, mil, net . . . ). The following sub-sections present some of these methods.

#### Public Namespace Windows

Referring back to FIG. 3, the X:\Internet\com directory contains only four entries: cnn, hp, ibm, and inca. These entries represent only a small subset of the registered members of the com domain. But, these entries are the only ones contained in this windowed view of the com public namespace.

DDS employs windowed views for all public namespace directories.

Windowed views, sometimes referred to as just windows or views, limit the scope of search requests to the "user's world". Both implicit and explicit methods are employed to construct views conforming to each user's interest profile.

DDS monitors user activities and creates interest profiles. Window entries are created implicitly from the user interest profiles and explicitly whenever a public namespace directory is accessed.

A user can easily determine his current windowed view by simply listing or displaying the contents of the appropriate public namespace directory. And, a user can easily add or remove entries by right clicking on the directory and selecting either an Add or Remove menu item.

A windowed view has the following properties:

1. A view is limited to a maximum number of entries (set by user or administrator).
2. An unsuccessful DDS\_CONNECT attempt to establish a connection to X:\Internet\com\sun (target sun not in the current window) creates a new entry in the following manner:
  - a. DNS (or equivalent) is queried for the target's IP address.
  - b. A DDS\_CONNECT (to the root of the sun domain tree) request is sent to IP address.
  - c. The target responds with the file handle and attributes of its domain tree root.
  - d. New entry (a directory named "sun") is created in the windowed view (X:\Internet\com\).
3. Window entries are maintained on a least recently used (LRU) and timeout basis.
4. A window entry may be pinned so that it will not be removed by the LRU or timeout mechanisms.

#### Domain Level Access Control

Limiting the scope of search to a windowed view drastically reduces the number of nodes participating in a search. However, a huge number of search requests are still likely to be delivered to well known domains.

As with other file system requests, executing a search request may require the user to have a particular authorization level. For example, a search request from a random user

received by boeing.com might be forwarded to all next level sub-domains, where it would be rejected by all sub-domains except one: the public.boeing.com sub-domain.

Another user, an American Airlines employee, submits a search request to boeing.com. This request, also forwarded to all sub-domains, is accepted by: aa.customer.boeing.com, marketing.boeing.com, and engineering.boeing.com in addition to public.boeing.com.

#### Search Request Routing

When a search request arrives at a domain portal, a domain manager may route the request on to specific sub-domains based on the user's identity, credentials and authorization. For example, when boeing.com received the search request from the American Airlines employee, a boeing.com domain manager might route the request on to only aa.customer.boeing.com, the American Airlines customer domain of the Boeing Company's global file system.

Selective routing filters out traffic that would otherwise be loading a domain's infrastructure. And, more importantly, selective routing actively directs the flow of incoming traffic and prevents malicious traffic from being indiscriminately sprayed into sub-domains.

The ability of a domain manager to actively control the flow of incoming traffic is consistent with a DDS principle (see U.S. Pat. No. 6,847,968 B2, identified earlier as being included in this disclosure) that a domain manager is in complete control of its resources. This principle applies to controlling the flow of network traffic within the domain.

#### Public Search Engines

An enterprise may decide for security (or other reasons) that its domain will not accept any network traffic from unknown sources. Searches will not be performed for anonymous users. However, there may be public object (document) collections scattered throughout the enterprise's domain, and these objects should be searchable and accessible by the public.

Top level enterprise domains may register the inverted files of public object collections with public search engines such as Google, Alta Vista, and Lycos.

#### Inverted File Forwarding

The registered inverted files may be periodically downloaded to public search engines and merged into composite inverted files, which are the files referenced to satisfy user queries. An individual query may be executed against multiple composite inverted files and the results may be merged into a single response.

The grouping that comprises a composite inverted file is generally an association of some sort. For example, all aviation related companies might be grouped together.

Individual composite inverted files are periodically reloaded and rebuilt, and then brought online, replacing the previous version composite inverted file. The update cycle period is days, weeks, or months depending on the particular group of inverted files.

When the results of a query against several composite inverted files are merged, each inverted file entry contains a global scope object ID that can be presented at any DDS portal to establish a connection to the object. Of course, the user must have proper credentials and authorization to actually establish a connection.

#### Inverted File Caching

Instead of downloading registered inverted files, the inverted files in remote top level enterprise domains may be remotely referenced by public search engines and dynamically cached as they're referenced. The cached indices may often constitute only a small fraction of the full set of

indices. When this is the case, it is more efficient to employ dynamic inverted file caching than to download and merge the complete inverted file.

#### Navigable Namespaces for Search Results—Parallel Namespaces

When a domain search is executed, a parallel namespace, populated with links to primary namespace objects satisfying the search criteria, is created and the file handle of the root of this namespace is returned to the process that initiated the search. The user (or application) may navigate this new namespace in the same manner used for the primary namespace. For example, a Windows PowerPoint user may double-click on the icon representing the !\_Search\_Results\_! directory within the \Internet\com\inca\ folder and the "(computer OR PDA) AND network" folder (and possibly other folders created by previous searches) comes into view. The user may continue double-clicking folders to explore this namespace and any of the content included within this space.

An individual link, which may be a directory, a symbolic link, a hard link, a global scope object ID, or even a URL, has the property that it points towards an object. There are two types of links:

1. Direct links—a direct link points at an object. A direct link is one reference removed from the target object.
2. Indirect links—an indirect link points toward an object or object collection. An indirect link points at a container object (such as a directory) containing links (either direct or indirect) to target objects. An indirect link never points directly at its target. It is always at least two references removed from its target.

Indirect links run parallel to primary namespace links.

For every directory link within a parallel search results namespace there exists a corresponding directory link in the primary namespace.

A search response may be categorized by the types of links it contains:

1. Direct response—the root of the parallel namespace contains only direct links. There are two forms of a direct response:
  - a. The object associated with the returned root file handle is a file containing a set of inverted index entries. The global scope object ID contained within each entry is a link to an object matching the query. This form, referred to as an un-scored direct response, returns a set of inverted index entries. The format of the response does not provide for the return of per entry relevance information. However, the return object is a file and therefore has an associated attribute structure that is included in the DDS\_CONNECT response. The attribute structure may be extended to convey the number of hits and/or a relevance score. This form associates a group score with a response.
  - b. The object associated with the returned root file handle is a directory. In this form, the textual names of the directory entries are derived from the global scope object IDs extracted from the set of inverted index entries comprising the search results.

The advantage of this form is that individual directory entries have associated attributes, and extensions to these attributes (see Attribute Extensions below) can incorporate the number of hits and/or a relevance score into the search results.

## 31

This form of direct response, referred to as a scored direct response, conveys additional information enabling the caller to discriminate between the results.

This form associates an individual score to each link within a response. Scored direct responses are usually sorted by relevance.

2. Indirect response—the root of the parallel namespace contains at least one indirect link. The object associated with the returned root file handle is a directory which may contain file objects and must contain at least one directory object. The significance of each is:
  - a. Each file object contains a collection of links, which may be scored or un-scored.
  - b. Each directory object is a scored link that enables the caller to discriminate between partial results contained within each directory.

#### Attribute Extensions

An object's attributes convey information about the object such as its size, its creation date, and its owner. Additional attributes may be defined that indicate how the object (or object collection) was scored by a search evaluation routine. By associating these scores with indirect links, the paths to the targets are, in effect, scored. This path scoring greatly facilitates a user's ability to quickly home in on the most relevant content.

#### Parallel Namespace Resolution

Indirect links are structural elements from which high resolution parallel namespaces may be constructed. The resolution of a parallel namespace is the degree to which the parallel namespace reflects its primary namespace. For example, suppose the eng domain 112's manager hoover returned a response containing links to one thousand objects. Consider the various forms the response might take:

1. The eng domain 112's manager hoover may use an un-scored direct response to return one thousand links. Then, it's up to the caller to peruse the associated objects and select the ones of interest.
2. The eng domain 112's manager hoover may use a scored direct response to return a collection of links sorted by relevance. Again, it's up to the caller to peruse and select. But, this time the caller has some hints as to which objects to peruse first.
3. The eng domain 112's manager hoover may employ an indirect response. A file handle for a directory object is returned. After the directory is read and the attributes for each entry fetched, the caller is able to discern that there are four partial responses (named pat, joe, bob, svrail) and that the bob domain 122 has the highest group score (both number of hits and relevance score). Although a scored direct response is used to report the content located in the pat domain 126, joe domain 124 and svrail domain 128, an indirect response is used to report the bob domain 122's findings. This response actually resides on the bob domain 122, which is where the search was executed.

Searching for the most relevant content, the caller navigates into the directory of the bob domain 122 and begins perusing its contents.

When there are only a few results to report, there is not much need to discriminate between results. The caller can just scan and evaluate each object. So, an un-scored direct response is an appropriate response form.

When there are a moderate number of hits to report, but they're all from the same source, a scored direct response is appropriate. Scores associated with each object convey hints as to which objects are the most likely to be relevant.

## 32

However, when the results number in the thousands (or millions!), search engines must structure the results in a manner that facilitates the caller's quest to locate some specific content. Indirect responses associate relevance scores with the paths to sub-domains, enabling the caller to discriminate between the results and navigate towards the desired content.

Referring again to FIG. 2, consider the processing of a search request targeting the inca domain 102:

1. The request, received by aztec (one of the two nodes hosting the inca domain 102), is forwarded on to the eng domain 112, the corp domain 116, the mrkt domain 118, and the sales domain 114.
2. The eng domain 112's manager hoover receives the request forwarded to the eng domain 112.
3. The eng domain 112's manager hoover forwards the request on to the bob domain 122 and svrail domain 128 and simultaneously initiates two local search threads which reference the cached indices of the pat domain 126 and joe domain 124.
4. The bob domain 122 receives a forwarded request and initiates a local search thread which references the (very up to date) inverted files maintained locally.
5. The svrail domain 128, a non-atomic domain, receives a forwarded request and forwards that request on to each of its sub-domains.
6. After a brief period the svrail domain 128 receives indirect responses from each sub-domain:
  - a. The svrail domain 128 composes its response by creating a response directory and then creating entries in that directory for each response received from a sub-domain.
  - b. Each directory created in the response directory has the same name as the sub-domain whose results are contained within.
  - c. The svrail domain 128 generates a composite relevance score and stores that data in the response directory's attributes.
  - d. Finally, the svrail domain 128 dispatches its response back to the eng domain 112's manager hoover.
7. In similar fashion, the eng domain 112's manager hoover creates a response directory and then creates entries in that directory for each of its sub-domain's responses. In this case, the joe and pat entries could be files (un-scored direct responses). The eng domain 112's manager hoover generates a composite relevance score and stores that data in the response directory's attributes. Finally, the eng domain 112's manager hoover dispatches its response back to aztec.
8. In similar fashion, aztec creates a response directory and then creates entries in that directory for each of its sub-domain's responses. However, searches of the corp domain 116 and of the mrkt domain 118 respectively reported "no hits", so no corp or mrkt entries were created in the response directory. aztec generates a composite relevance score and stores that data in the response directory's attributes. Finally, aztec dispatches its response back to the caller.
9. The caller receives the response from aztec, which contains:
  - a. a SUCCESS status indicator,
  - b. a file handle, and
  - c. attributes containing a relevance score and an indicator as to the type of object associated with the returned file handle (file or directory).

10. The caller reads the response directory and discovers the eng and sales responses. Further perusal discloses that the results namespace hierarchy reflects the primary namespace.

#### Namespace Discrimination

Any attempt to create a directory within a search results parallel namespace is interpreted (and processed!) as a search request. The directory name (the last component of the CreateDirectory pathname parameter) is interpreted both as a search specification string and as the new directory's name. So, for example, if the user navigates to the folder "\Internet\com\inca!\\_Search\_Results\_!(computer OR PDA) AND network\eng\" and attempts to create a directory, the search mechanism is invoked to search the eng domain 112 with the directory name interpreted as the search specification. The content searched includes only the content contained within the current parallel namespace (the results of a previous search).

This is a means of selectively refining a search by navigating into a region of dense results and then applying additional search criteria to further discriminate the results. Note that various dense regions of a search may be selectively discriminated using different criteria.

#### Namespace Coherency

The "\...!\\_Search\_Results\_!(computer OR PDA) AND network\eng\" folder contains content matching the query "(computer OR PDA) AND network\" within the eng domain 112. This folder may be searched for content matching "Palm OR Handspring" and the results (references to Palm or Handspring PDAs with networking capabilities) will be stored the "\...!\\_Search\_Results\_!(computer OR PDA) AND network\eng\Palm OR Handspring\" folder.

Namespace coherency is the namespace property exhibited by the example just presented. Each component of the pathname of the "Palm OR Handspring" results folder accurately describes the content of that folder:

1. (computer OR PDA) AND network\  
inca content matching the query.
2. (computer OR PDA) AND network\eng\  
eng (an inca sub-domain) content matching the query.
3. (computer OR PDA) AND network\eng\Palm OR  
Handspring\  
eng content matching the query:  
((computer OR PDA) AND network) AND (Palm OR  
Handspring)

#### A Distributed Out-of-Band Signaling Mechanism

Geographically distributed applications may employ a global file system (such as DDS) to concurrently access a common file and, through a series of read and write operations, interact with each other. The DDS consistency mechanism guarantees clients that a read request always returns the most recently written data, even when continents and oceans separate the reader and the writer. Obviously, a strong consistency guarantee makes shared file access a more viable form of interprocess communication.

Without strong consistency guarantees, the burden of clear and consistent communications must be carried by the applications. This, of course, adds complexity to all applications, and requires that application level programmers provide a solution to a difficult problem that, most likely, lies outside of their areas of expertise.

A file server and its client-side components may be viewed as a distributed process. The file service collectively provided is enhanced by the out-of-band consistency control operations employed to maintain the coherency of file data cached by client-side components.

In currently deployed distributed file systems there are basically two private messages (out-of-band consistency messages) used to maintain client-side cache consistency:

1. Invalidate—discard cached file data (file identifier included in the message).
2. Recall—flush file modifications to server and then discard cached file data (file identifier in the message).

These messages are private, flowing only between a server and its client-side components. They are also out-of-band, meaning these messages exchange control information outside of normal file traffic data flow. An application process accessing a remote file is usually unaware of whatever consistency operations are being performed between the file server and its client-side components. In fact, the Unix, Linux and Windows operating systems do not provide any mechanism that enables an application to monitor (or be aware of) client-side cache consistency operations.

#### Out-of-band Signaling between Distributed Processes

Although a distributed consistency mechanism for maintaining file system cache coherency only requires a few message types (Recall, Invalidate, and perhaps a few variants), this type of code is difficult to develop and thoroughly debug. So, once the effort has been made to implement and debug an out-of-band communication channel the question arises: Can other cooperating distributed processes benefit from out-of-band control channels? The answer is yes, and the following three scenarios illustrate how out-of-band channels can provide a superior interprocess communications infrastructure:

#### 1. Two Very Remote Processes Communicating Via an Intermediary File Server:

- a. Both processes open the same file in write append mode and both request out-of-band notifications whenever the other process writes to the file.
- b. Process A "sends" a request to Process B by executing a synchronous write to the shared file.
- c. The intermediary file server receives the write data, appends it to the end of the file, issues a notification to Process B, and dispatches a response (to the write request) back to Process A.
- d. Process B receives the notification and immediately reads the shared file (from its current access position to end of file) and thereby receives the request from Process A.
- e. Process B performs the operation requested by Process A and synchronously appends its response to the end of the shared file.
- f. The intermediary file server receives the write data, appends it to the end of the file, issues a notification to Process A, and dispatches a response back to Process B.
- g. Process A receives the notification and immediately reads the shared file (from its current access position to end of file) and thereby receives the response from Process B.
- h. Process A continues sending requests and receiving responses until it decides to terminate the session.
- i. After session termination, the shared file is a complete log of the session. It may be saved or discarded according to established policies.

#### 2. Two Processes Communicate Via a Memory Mapped File:

- a. Both processes map the same file into their address space and both request out-of-band notifications whenever the other process modifies a specified location in the shared address space.
- b. Process A constructs a request to Process B in the shared memory and then synchronously writes the

35

address of the request to the location specified by Process B. The synchronous write flushes all modifications to the server.

- c. The intermediary file server receives the flush request (the modifications), recognizes that Process B's specified location has been modified and issues a notification to Process B, and dispatches a flush response back to Process A.
  - d. Process B receives the notification and immediately reads its specified location and retrieves a pointer to the request from Process A.
  - e. Process B performs the operation requested by Process A and constructs its response in the shared memory and then synchronously writes the address of the response to the location specified by Process A. The synchronous write flushes all modifications to the server.
  - f. The intermediary file server receives the flush request (the modifications), recognizes that Process A's specified location has been modified and issues a notification to Process A, and dispatches a flush response back to Process B.
  - g. Process A receives the notification and immediately reads its specified location and retrieves a pointer to the response from Process B.
  - h. Process A continues sending requests and receiving responses until it decides to terminate the session.
3. Real-Time Video Broadcasting Via a Distributed File System Object:

- a. At a possibly pre-established time, a broadcasting source creates a channel object or source object within a distributed file system and begins writing an audio and/or video data stream to the file system object.
- b. Each viewer (viewing client application) establishes a connection to the distributed file system channel object and receives from the distributed file system an object identifier in the form of a file descriptor. Each viewer then registers to receive update notifications whenever the channel object (identified by the object identifier parameter contained in the registration request) is modified.

An update notification is a notification that the channel object has been modified which also includes all data necessary to update any local images of the channel object buffered or cached at other nodes or sites to the same state as the remote channel object located at the broadcasting source node. The update data may be compressed (AVI, MPEG, MP4, DVD, FLV, MOV, . . . ) or uncompressed.

- c. There may be thousands or millions of active viewers. A distributed file system, providing the infrastructure connecting all viewers to the broadcasting channel object, includes a hierarchy of intermediate nodes (distributed file system nodes that are not the broadcasting source node), each node usually incorporating a file system proxy caching capability. Intermediate nodes provide the interconnection framework that connects the viewer nodes (nodes that are supporting at least one viewer application) to the broadcasting source node. A node can simultaneously be both a viewer node and an intermediate node. "Next level node" is a term used to refer to a node that is one level further removed from the broadcasting node. It can be either a viewer node or an intermediate node.

When an intermediate node receives its first update notification registration request for an identified distributed file system channel object, the node registers with a downstream node (closer to the source object,

36

and possibly the source broadcasting node) to receive update notifications. The intermediate node also records the address of the requesting node or the identifier (pid) of the requesting viewer process so that future update notifications, when received, can be forwarded on towards all viewer processes that are upstream of the intermediate node or are executing on the intermediate node.

When an intermediate node receives additional update notification registration requests (from other applications or upstream nodes closer to the viewing applications) for a specific file system channel object, the intermediate node only needs to record the address of the requesting node or the identifier (pid) of the requesting process. The intermediate node does not need to register with the downstream node to receive update notifications since it has already registered to receive update notifications.

- d. At the beginning of a broadcast, the video source may write the first video frame (the title frame) and then not update or change that frame for a period of time which might extend to ten minutes or more. This is the "tune-in period" during which viewing applications establish connections to the channel object and then register to receive update notifications. The title frame is transferred to a viewer when the connection is initially established and then no additional network traffic is required until the video source begins writing new data to the channel object.
- e. At some point the video source begins writing to the channel object located at the broadcasting source node (probably at a regular update rate such as thirty times per second). Each write (frame update) modifies the channel object, which causes the broadcasting node to generate an update notification message. The update notification message is then multicast (a single message is simultaneously transmitted to multiple network nodes) to all registered remote viewers that had registered with this node (the broadcasting node). It is possible, but unlikely, that one or more registered viewers are local viewing applications executing on the broadcasting node. The channel object at the broadcasting node has already been updated, but each local viewer receives an update notification so that the viewing application can become aware of the update and perform any processing that might be required.
- f. When an upstream node (again, towards the remote viewers) receives an update notification, it updates its copy of the channel object, which causes the upstream node to generate its own update notification message. This update notification message is then multicast to all nodes further upstream that have registered with this node (the upstream node). If there are any local viewing applications executing on the upstream node, each local viewer receives an update notification so that the viewing application can become aware of the update and perform any processing that might be required.

- g. Nodes maintaining an image of the channel object (viewing nodes, intermediate nodes and the broadcasting node) are usually configured with a file caching capability. This capability enables a node to operate in a manner very similar to a digital video recorder (dvr) and provides a viewing node with the ability to rewind and replay a live (or canned) video stream. However, the file caching capability's most significant contribution to real-time network based broadcasting is that it provides viewing nodes with the ability to read ahead



37

of the data being presented on the viewer. This allows the viewing application to deliver high quality presentations even when the network is delivering update notifications on a less than timely basis. Finally, caching allows a program to be saved and replayed long after it has been broadcast without burdening the downstream nodes.

- h. After audio and video broadcasting via the Internet has become well established, there will be millions of channels vying for an individual's selection. From an individual's perspective, there will be an overwhelming number of channels. However, each channel may be tagged with textual descriptors which can be used as index terms for the distributed file system's integrated indexing and search mechanism, which is also disclosed in this patent application. Individuals will therefore be able to invoke the capabilities of a powerful search mechanism to help them locate the channels of most interest.

Three methods illustrating the use of an out-of-band communication channel were presented above, but there are practically an unlimited number of ways that multiple processes can coordinate their activities using out-of-band communications.

Out-of-band communication channels may be enhanced by extending the functionality of the consistency callback mechanisms employed by local file systems such as NTFS and distributed file systems such as DDS, NFS and CIFS in the following manner:

1. Provide a method for event notification registration, by which application programs can register with the file system the file events for which notifications should be delivered. The preferred means of performing this registration is to define an IOCTL (I/O Control) function that identifies a specific file, directory, or object ("file" for the remainder of this disclosure) and then specifies the associated notification trigger event(s).

The Linux man page SYNOPSIS for IOCTL is:

```
int ioctl(int fd, int cmd, char *argp);
    fd—the file descriptor of the file to monitor
    cmd—the type of request
    argp—arguments pointer
```

A registration specifies either synchronous or asynchronous notifications:

- a. cmd=SYNCNOTIFY

The IOCTL system call returns when a delivered notification message matches the file events specified in this registration.

- b. cmd=ASYNCTIFY

The IOCTL system call returns immediately and the notification routine whose address is specified as the Notification Routine Address parameter (described below) is called whenever a delivered notification message matches the file events specified in this registration.

The arguments pointer contains the address of an event notification descriptor. In an embodiment illustrated in FIG. 4, an event notification descriptor contains six fields:

- a. Event—specifies the event(s) which will generate a notification:
  - ACCESS\_READ—a read access generates a notification,
  - ACCESS\_WRITE—a write access generates a notification,
  - ACCESS\_RW—either read or write access generates a notification,

38

ACCESS\_OPEN—file open operation generates a notification,

ACCESS\_CLOSE—file close operation generates a notification,

SEND\_UPDATE—specifies that update data should be included in notifications.

- b. Event Range Start—This specifies the starting address of a byte range (the event range) within the file. File accesses outside of the event range will not generate a notification.

- c. Event Range End—This specifies the ending address of the event range.

- d. Update Range Start—The start of an update range within which modified file data is packaged and delivered as an integral part of the notification message. This allows a WRITE\_ACCESS notification to also include the new file data required to update a cached image of the file.

- e. Update Range End—The end of the update range.

- E Notification Routine Address—The address of the application's notification routine to be called when asynchronous notifications are being employed. This field is ignored when synchronous notifications are being used.

2. Store a file's associated notification descriptor(s) as extended attributes. These extended attributes, called notification attributes, are, in most respects, handled in the same manner as the regular attributes:

- a. the notification attributes are cached and valid wherever the regular attributes are cached and valid,

- b. the notification attributes are under the same consistency controls as the file's data and metadata (attributes),

- c. the notification attributes are present (at least momentarily) and valid and may be referenced at some point in the processing of every file access request targeting the associated file or directory,

- d. the notification attributes will persist across a restart (or reboot) of a failed origin server or intermediate caching node;

- e. however, they do not persist beyond the last connected application or process terminating its connection to (closing) the file.

3. Examine each file access request and detect when the request satisfies the conditions of a notification trigger event described in the file's notification attributes.

4. When a notification is triggered, notify all registered sites where the file is currently being accessed by extending the capability of DDS's (or other) distributed consistency control mechanism such that it can deliver notification messages in addition to RECALL and DISABLE messages. A notification message includes a notification message descriptor, one embodiment of which is illustrated in FIG. 5, which conveys the trigger(s) that initiated the dispatch of the message. In an embodiment illustrated in FIG. 5, a notification message contains six fields:

- a. Event—specifies the event which generated the notification:
  - FILE\_WAS\_READ,
  - FILE\_WAS\_WRITTEN,
  - FILE\_WAS\_OPENED,
  - FILE\_WAS\_CLOSED,
  - UPDATE\_INCLUDED—notification message includes update data.

39

- b. Event Range Start—This specifies the starting address of a byte range (the event range) within the file that generated the notification.
- c. Event Range End—This specifies the ending address of the event range. 5
- d. New\_Data\_Start—The starting address of the new file data contained within the notification message required to update a cached image of the file.
- e. New\_Data\_End—The ending address of the new file data. 10
- f. Site\_Identifier—Identifies the site that performed the file access that triggered the notification.
- 5. Each registered site compares the trigger(s) conveyed in a delivered notification message with the file events registered for the file at this site. 15
- 6. When the trigger(s) conveyed in a delivered notification message match the file events registered for the file at a registered site, deliver the notification to the application: 20
- a. Synchronous notification registration—Return from IOCTL call.
- b. Asynchronous notification registration—Invoke application's notification routine. 25

In the case where the consistency mechanism is used for broadcasting content, such as video, the notification attributes and trigger event will usually be limited to file updates, i.e., when data is written to the file. The notification routine will include an instruction to send the updated data along with the notification of the update. Similarly, the notification message descriptor will thus include the updated data. See FIGS. 4 and 5. 30

Although the present invention has been described in terms of the presently preferred embodiment, it is to be understood that such disclosure is purely illustrative and is not to be interpreted as limiting. The methods disclosed herein integrate a document retrieval mechanism into a distributed file system. 35

The preferred retrieval mechanism is a full text retrieval system. However, the disclosed methods are also applicable to other retrieval systems which may be woven into the fabric of distributed file systems using the disclosed methods. Though described in the context of DDS virtual file servers, the disclosed methods are also applicable to other distributed file systems such as NFS, CIFS and Appletalk, and to local file systems such as NTFS, UFS, and EXT2FS. 40

Consequently, without departing from the spirit and scope of the disclosure, various alterations, modifications, and/or alternative applications will, no doubt, be suggested to those skilled in the art after having read the preceding disclosure. Accordingly, it is intended that the following claims be interpreted as encompassing all alterations, modifications, or alternative applications as fall within the true spirit and scope of the disclosure including equivalents thereof. 45

What is claimed is:

1. A method of communicating various file access read and write requests performed on a file located at an origin file server site within a distributed file system that includes multiple caching node sites to a plurality of applications running at different sites in the distributed file system, comprising: 60

an origin file server site within a distributed file system: receiving from each of a plurality of applications running at different sites in the distributed file system: 65

40

a request for notification whenever a file access request received at any site in the distributed file system triggers a notification, the request for notification containing:

a file descriptor identifying the file for which the application requests notification of file access requests; and,

a specification of the type of file access activity required to trigger a notification;

for the file identified by the file descriptor:

creating a notification attribute as an extended attribute for the file;

storing the request for notification within the notification attribute; and

employing a consistency mechanism of the distributed file system to ensure all images of the notification attribute at all sites currently accessing the file are the same; and

when any distributed file system site receives a request to access the file:

referencing each of the notification attributes to determine which of the different sites in the distributed file system at which the file is currently being accessed have requested a notification message; and

delivering a notification message to each of the different sites in the distributed file system that have requested a notification message.

2. The method of claim 1 wherein registering a request for notification further comprises an IOCTL system call function having arguments which include the file descriptor.

3. The method of claim 2, wherein the IOCTL system call function returns when a notification has been triggered.

4. The method of claim 2, wherein the file descriptor includes a notification routine address.

5. The method of claim 4, wherein the IOCTL system call function returns immediately and the notification routine identified by the notification routine address is called when a notification is received.

6. The method of claim 2, wherein the IOCTL system call function includes a description of the type of file access required to trigger a notification.

7. The method of claim 1 further comprising:

discarding the notification attribute of the file when all of the applications that have registered for that specific notification have terminated their connections to the identified file.

8. The method of claim 1 wherein delivering a notification message to each of the different sites in the distributed file system that have requested a notification message further comprises employing IP multicast to send the notification message to all of the sites.

9. The method of claim 1 wherein the request for notification includes at least one of:

an identification of the event which will trigger the request for notification;

the event range of the access which will trigger the request for notification;

the update range of the access which will trigger the request for notification; or

the notification routine address to be called when asynchronous notifications are being employed.

10. The method of claim 1 wherein the identification of the event includes at least one of:

an identification of the event which triggered the notification;

## 41

the event range of the access which triggered the notification;  
 the update range of the access which triggered the notification; or  
 the notification routine address to be called when asynchronous notifications are being employed.

11. The method of claim 1 wherein the notification message includes at least one of:

an identification of the event which triggered the notification message;  
 the event range of the access which triggered the notification message;  
 the update range of the access which triggered the notification message; or  
 a site identifier which identifies the site that performed the file access which triggered the notification message.

12. The method of claim 1 further comprising:  
 the file access request that triggered a notification is not processed by the distributed file system until all sites receiving the notification have acknowledged receipt of the notification.

13. The method of claim 1, wherein creating a notification attribute as an extended attribute for the file includes:

## 42

the notification attributes are cached and valid wherever the regular attributes are cached and valid.

14. The method of claim 1, wherein creating a notification attribute as an extended attribute for the file includes:  
 the notification attributes are under the same consistency controls as the file's data and metadata.

15. The method of claim 1, wherein creating a notification attribute as an extended attribute for the file includes:  
 the notification attributes are present and valid and may be referenced at some point in the processing of every file access request targeting the associated file or directory.

16. The method of claim 1, wherein creating a notification attribute as an extended attribute for the file includes:  
 the notification attributes will persist across a restart of a failed origin file server or intermediate caching node.

17. The method of claim 1, wherein creating a notification attribute as an extended attribute for the file includes:  
 the notification attribute will not persist beyond the last connected application or process terminating its connection to the file.

\* \* \* \* \*